Unit 1

Exploring and interpreting data

Introduction

We live in an age of data! In recent years an information explosion has revolutionised our whole environment. With the development of high-speed and high-capacity computers and related technological advances, vast amounts of data are being generated every minute. Information pours in from the media, government agencies, researchers, commercial companies and a host of other sources, and we must learn to make rational choices based on some kind of summary and analysis of this information. The aim of this module is to give you some tools for making sense of data.

If you have already studied an introductory module on statistics, you will have learned about some techniques for performing relevant tasks. The aim of this module is to build on this prior study. This will be achieved by investigating a greater range of techniques as well as by providing a deeper understanding of the techniques that you might have already come across.

We begin here, in Unit 1, by revising techniques for 'getting a feel' for data. Data exhibit variation – so the characteristics whose values change from one individual to another are called variables – and we are initially interested in exploring informally how values of the data vary as well as possible relations between variables. This is done by drawing pictures of the data and by summarising the data numerically. Although getting a feel for the data is likely to involve techniques you have already met, its continued importance is not to be underestimated. Insights gained at this stage guide the rest of the analysis by, for example, highlighting which techniques are appropriate (or, more importantly, which are *not* appropriate) and providing a guide as to whether whatever conclusions are reached appear to be reasonable.

To get started, in Section 1, we introduce some data, and a few of the important words used to describe them. Then, in Section 2, we explore the question of whether a dataset can be thought of as representing an entire population or as just a sample from it. We describe some plots that can be used to depict data in Section 3. Interest is in the *distribution* of the data, that is, in the pattern by which data values vary: a distribution has attributes such as shape, location and spread. Numerical quantities that can be used to summarise aspects of a distribution of data are then described in Section 4. Finally, Section 5 returns to the graphical theme, this time using plots to compare variables and to explore relationships between them.

An important tool in modern statistics is a statistical software package which is used to get your computer to deal with the number-crunching drudgery underlying most statistical analysis. This tool allows statisticians to focus on selecting the right techniques for the right situation – and on interpreting the results. The specific statistics package that will be used in this module is Minitab. The basic use of Minitab, and its implementation of the techniques revised in this unit, will be covered in stages as you work

through the unit. However, it is not necessary to complete the computer work associated with one section before moving on to the next section. So study of the sections involving Minitab can, if you wish, be deferred to the end of your study of this unit.

1 Data

For data analysis we need data. So here we introduce some datasets to explore. As you will see, the datasets described here are mostly small. This is deliberate, so that you can more easily see what is going on and have the opportunity to consider individual data points. However, it is worth bearing in mind that many datasets are large enough that inspecting all the values individually becomes tedious. Yet other datasets are so large that inspecting more than a tiny minority of data points becomes completely impractical.

Example 1 *Nuclear power stations*

Grafenrheinfeld nuclear power plant, Bavaria, Germany

The first dataset is a very simple one. Table 1 shows the number of operational nuclear power stations in various countries throughout the world in 2014.

Table 1 Operational nuclear power stations

Country	Number	Country	Number
Argentina	3	Mexico	2
Armenia	1	Netherlands	1
Belgium	7	Pakistan	3
Brazil	2	Romania	2
Bulgaria	2	Russia	34
Canada	19	Slovakia	4
China	23	Slovenia	1
Czech Republic	6	South Africa	2
Finland	4	South Korea	23
France	58	Spain	7
Germany	9	Sweden	10
Hungary	4	Switzerland	5
India	21	UK	16
Iran	1	Ukraine	15
Japan	48	USA	99

(Source: International Atomic Energy Agency, https://www.iaea.org)

Notice that in Table 1 each country has at least one operational nuclear power station. So the dataset includes only those countries that had operational nuclear powers stations in 2014 – and not any countries that did not.

Questions to which these data are relevant include: if a country has operational nuclear power stations, how many does it typically tend to have? and, how many countries have lots of operational nuclear power stations?

Example 2 *UK workforce*

Our second dataset, which we give in Table 2, gives the total workforce size in the UK during the last quarter of 2015, categorised into different occupation types. Also given for each occupation type are the numbers employed, broken down by gender.

Which occupation type is this female working in?

Table 2 Composition of the UK workforce (in millions) in the last quarter of 2015

Occupation type	Male	Female	Total
Managers, directors & senior officials (Managers)	2.118	1.153	3.271
Professional occupations (Professional)	3.172	3.014	6.186
Associate professional & technical (Technical)	2.466	1.901	4.367
Administrative & secretarial (Administrative)	0.860	2.480	3.340
Skilled trades	3.107	0.334	3.441
Caring, leisure & other services (Caring & leisure)	0.520	2.390	2.910
Sales & customer services (Sales)	0.918	1.568	2.486
Process, plant & machine operatives (Operatives)	1.743	0.230	1.973
Elementary occupations (Elementary)	1.852	1.565	3.417

(Source: Office for National Statistics, https://www.ons.gov.uk)

Questions which these data might be used to answer include: is the workforce evenly spread across the occupation types? and, in which occupation types (if any) is there a gender imbalance?

Example 3 *Number of children*

Many of the datasets considered in M248, like those in Examples 1 and 2, are of direct contemporary interest. Other datasets are older but retain their interest for historical reasons. One such can be found in Table 3. These data were sampled from the 1941 Canadian census and comprise the numbers of children born to Protestant mothers in Ontario who were then aged 45–54 and had been married aged 15–19. In addition, the data in Table 3 are confined to mothers who had been educated for seven years or more. The number of entries in the dataset is 35.

Table 3 Number of children

0	4	0	2	3	3	0	4	7	1	9	4	3	2	3	2	16	6	0	13	6	6	5	9	10	5	4	3	3	5	2	3	5	15	5

Here, we might ask: what is the distribution of family size for mothers in the category described? and, do they typically have small or large families?

Example 4 *Membership of sports clubs*

Table 4 gives the percentages of adults (aged 16+) who were members of sports clubs in 2014–15 in the 49 'sport partnership areas' of England. (These 49 areas covered the whole of England.)

Table 4 Percentages of adults in sport partnership areas who were members of sports clubs

16.8	20.7	22.3	20.3	19.3	24.0	27.7
21.9	19.7	21.2	20.9	22.2	20.4	22.9
24.0	19.9	22.6	19.1	19.9	23.8	20.4
16.7	22.3	22.3	19.1	23.7	18.6	21.9
24.9	24.0	25.0	25.4	21.8	18.9	23.5
21.3	17.9	17.4	21.8	22.3	24.3	21.2
22.5	22.8	22.3	27.6	20.4	23.1	19.9

(Source: Sport England, http://www.sportengland.org)

Questions to which these data are relevant include: typically what percentage of adults in an area were members of a sports club? and, were there are any areas where sports club membership was particularly low or particularly high?

Example 5 *Response inhibition training*

Cartoon used by authors to illustrate their study

Table 5 contains data about 10 of the 83 participants in a clinical trial investigating a treatment to help with weight loss: response inhibition training. It was hoped that response inhibition training would help reduce the consumption of high-energy foods by training people to react less favourably to pictures of such foods. In the trial, the participants were either in the treatment group (labelled 'T') receiving response inhibition training, or in a control group (labelled 'C') doing similar activities to the response inhibition training, but not linked to food. The number of training sessions the participant took part in, and the weight change (in kg) over the first two weeks is also given. (A negative weight change is a weight loss, a positive weight change is a weight gain.)

Table 5 A selection of data from a weight loss trial

Treatment group	T	T	T	C	T	T	C	C	C	T
Number of sessions	4	4	4	4	3	4	4	4	4	4
Gender	F	F	F	F	F	F	M	M	F	F
Weight change	−0.6	1.0	0.2	0.8	−1.9	−1.3	0.2	1.4	0.2	−1.5

(Source: Lawrence, N.S. et al. (2015) 'Training response inhibition to food is associated with weight loss and reduced energy intake', *Appetite*, vol. 95, pp. 17–28)

With such data, the key question is whether weight loss is greater in the treatment group compared with the control group.

Example 6 *Surgical removal of tattoos*

Table 6 contains clinical data from 55 patients who had forearm tattoos removed. Two different surgical methods were used; these are denoted by A and B in the table. The tattoos were of large, medium or small size, either deep or at moderate depth. The final result is scored on an integer scale from '1' to '4', where '1' represents a poor removal and '4' represents an excellent result. The gender of the patient is also shown.

Here, with these data, interesting questions include: is one method more likely to produce a better result than the other?, does the quality of the result depend on the size of the tattoo? and, does the quality of the result depend on the depth of the tattoo?

Modern laser tattoo removal

Table 6 Surgical removal of tattoos

Method	Gender	Size	Depth	Score	Method	Gender	Size	Depth	Score
A	M	large	deep	1	B	M	medium	moderate	2
A	M	large	moderate	1	B	M	large	moderate	1
B	F	small	deep	1	A	M	medium	deep	2
B	M	small	moderate	4	B	M	large	deep	3
B	F	large	deep	3	A	F	large	moderate	1
B	M	medium	moderate	4	B	F	medium	deep	2
B	M	medium	deep	4	A	F	medium	deep	1
A	M	large	deep	1	A	M	medium	moderate	3
A	M	large	moderate	4	B	M	large	moderate	3
A	M	small	moderate	4	A	M	medium	deep	1
A	M	large	deep	1	A	F	small	deep	2
A	M	large	moderate	4	A	M	large	moderate	2
A	F	small	moderate	3	B	M	large	deep	2
B	M	large	deep	3	B	M	medium	moderate	4
B	M	large	deep	2	B	M	medium	deep	1
B	F	medium	moderate	2	B	F	medium	moderate	3
B	M	large	deep	1	B	M	large	moderate	2
B	F	medium	deep	1	B	M	large	moderate	2
B	F	small	moderate	3	B	M	large	moderate	4
A	F	small	moderate	4	B	M	small	deep	4
B	M	large	deep	2	B	M	large	moderate	3
A	M	medium	moderate	4	B	M	large	deep	2
B	M	large	deep	4	A	M	large	deep	3
B	M	large	moderate	4	A	M	large	moderate	4
A	M	large	deep	4	B	M	large	deep	2
B	M	medium	moderate	3	B	M	medium	deep	1
A	M	large	deep	1	A	M	small	deep	2
B	M	large	moderate	4					

(Source: Lunn, A.D. and McNeil, D.R. (1988) *The SPIDA Manual*, Statistical Computing Laboratory, Sydney)

Activity 1 *Comparing the structure of data*

In the preceding examples, six different datasets have been provided. How are the structures of these datasets similar, and how are the structures different?

As you have seen in Activity 1, the different datasets have different structures. In statistics, the objects or individuals in a dataset are known as **observations**, **cases** or **sampling units**. The characteristics are referred to as **variables**. And the pattern of variation in the values of a variable is called their **distribution**.

Variables can be categorised into different types; a variable's type is determined by what characteristic it is representing.

The 'interval' might be of infinite length.

Continuous variables correspond to numerical characteristics where any value within an interval of values is possible. For example, they often correspond to measurements such as weight, length or temperature.

For counts, only zero and positive integer values are possible.

Discrete variables correspond to numerical characteristics where only particular values are possible. For example, they often correspond to counts, such as of the number of children in a family or of the number of bus routes in a city.

In M248, we will think of categorical variables as being distinct from continuous and discrete variables, which are numerical variables. Categorical variables are, however, quite often considered to be a special type of discrete variable.

For both continuous and discrete variables, all the possible values are numbers. However, for **categorical variables**, values are just labels – they indicate which one of a number of groups an observation belongs to. An example is a variable that indicates whether someone never smoked, is an ex-smoker or is a current smoker. Note that the labels attached to each group might be numerical, such as '1', '2' and '3', or might not be numerical, such as 'A', 'B' and 'C'.

Categorical data can be further split into two types: nominal and ordinal. For **ordinal (categorical) data**, the categories have a natural ordering. For example, a response to a question on a five-point scale which represents 'strongly disagree', 'disagree', 'neither agree nor disagree', 'agree' and 'strongly agree' is ordinal. If there is no natural ordering, the categorical data are said to be **nominal**, for example, when the possible categories are types of woodland habitat, perhaps 'native pinewoods', 'native lowland woodland', 'plantations', 'ancient woodland' and others.

Activity 2 *Categorising variables*

For each of the following variables, state whether it is continuous, discrete, nominal or ordinal.

(a) In Table 1, the number of operational power stations.

(b) In Table 4, the percentage of adults who were members of sports clubs.

(c) In Table 6, the surgical method used.

(d) In Table 6, the size of the tattoo.

(e) In Table 6, the final result of the tattoo removal.

(f) In Table 2, the total number of people employed.

We will return to the aspects of the above categorisation of variables that are of most importance in this module in Subsection 2.1 of Unit 2.

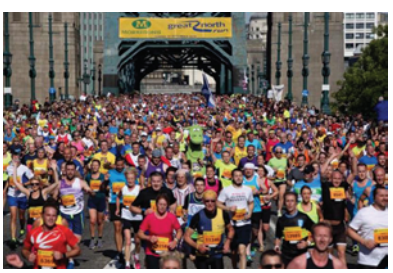
These include the point made in the solution to Activity 2(f).

Recall that in Tables 2, 5 and 6, the data consisted of two or more variables measured on the same objects. In each case, the groups of variables are known as **linked variables**. For example, in Table 5, the variables corresponding to the treatment group, the number of training sessions attended, gender and weight change are linked. This is because the values of all four variables are recorded for each object (in this case, for each participant in the study).

Activity 3 *Runners*

The dataset given in Table 7 relates to 22 competitors in the Great North Run. Blood samples were taken from 11 runners before and after the run, and from another 11 runners who collapsed near the end of the race. The measurements are plasma β endorphin concentrations in picomoles per litre (pmol/l). Unless you have had medical training, you are unlikely to know precisely what constitutes a plasma β endorphin concentration, or what the units of measurement mean. This is a common experience among statisticians when working with data from specialist experiments. What matters most here is that some physical attribute can be measured, and that the measured value is important to the experimenter. The statistician is prepared to accept that running may have an effect on the blood, and will ask for clarification of medical questions as and when the need arises.

β is the Greek lower-case letter beta, pronounced 'beeta'.



The Great North Run is the largest half-marathon in the world. It has been run annually since 1981, nowadays attracting over 50 000 competitors. Here, runners stream over Newcastle's Tyne Bridge in the 2015 event.

Table 7 Blood plasma β endorphin concentrations (pmol/l)

Normal runner before race	Same runner after race	Collapsed runner
4.3	29.6	66
4.6	25.1	72
5.2	15.5	79
5.2	29.6	84
6.6	24.1	102
7.2	37.8	110
8.4	20.2	123
9.0	21.9	144
10.4	14.2	162
14.0	34.6	169
17.8	46.2	414

(Source: Dale, G. et al. (1987) 'Beta-endorphin: a factor in "fun run" collapse?', *British Medical Journal*, vol. 294, p. 1004)

In this dataset, there are three variables which correspond to the three columns of the table.

(a) For each variable, state whether it appears to be continuous or discrete.

(b) Are any of the variables linked? If so, which?

This section has focused on thinking about tables of data, and the contexts in which the data arose. A summary of the terminology is as follows.

When describing a dataset, the following terminology is used.

- **Observations** (or **cases**, or **sampling units**) refer to objects (people, countries, ...) on which characteristics are recorded.
- **Variables** are the characteristics recorded, and the pattern of variation of a variable is its **distribution**.
- Variables are **linked** if they are each recorded for the same observations.
- A variable is **continuous** if its values are numerical and all values in an interval are possible.
- A variable is **discrete** if its values are numerical but only particular values (typically, integers) are possible.
- A variable is **categorical** if its values indicate to which group an observation belongs.
- A categorical variable is **ordinal** if its values correspond to labels which have a natural ordering.
- A categorical variable is **nominal** if its values correspond to labels but the labels do not have a natural ordering.

Even though you have not yet drawn your first graph of any data, or calculated your first numerical summary, by thinking about the data structure you are taking your first important steps in analysing the data. As you will see as you study the module, knowing what types of variables you have, and whether or not they are linked, will dictate which statistical techniques can reasonably be applied.

Exercises on Section 1

Exercise 1 *Types of variables in the Crime Survey for England and Wales*

The Crime Survey for England and Wales (CSEW) aims to capture people's experience of crime. It is a large survey run by the Office for National Statistics to help the UK government understand the true level of crime in the country. In the survey, many variables are recorded, some of which are given below. For each of these variables, state whether it appears to be a nominal, ordinal, discrete or continuous variable.

(a) The respondent's age, given in years.

(b) The length of time the respondent has been living in their current area (in years), grouped as '< 1', '1–2', '2–3', '3–5', '5–10', '10–20' or '20+'.

Respondents with lengths of time of exactly 2, 3, 5, 10 or 20 years were put in the relevant group along with longer times; e.g. a respondent who has lived in the area for 3 years was assigned to the '3–5' group.

(c) The marital status of the respondent, coded as 'single', 'married', 'separated', 'divorced', 'widowed' or 'civil partnership'. (This is a slight simplification of the categories used.)

(d) How safe the respondent feels walking alone in their local area after dark, coded as 'Very safe', 'Fairly safe', 'A bit unsafe' or 'Very unsafe'.

Exercise 2 *Linking of variables in the CSEW*

(a) In the CSEW for 2014–15, would the two variables 'how safe the respondent feels' and 'age in years' be linked or not linked?

(b) The CSEW for 2015–16 surveyed different people to those surveyed in 2014–15. Are the variables 'how safe the respondent feels' in the 2014–15 survey and 'how safe the respondent feels' in the 2015–16 survey linked or not linked?

2 Populations and samples

In Section 1, various datasets were introduced, for example, about nuclear power stations and about the removal of tattoos. Further, in that section you learned that the datasets could be thought of as consisting of a set of objects (observations) about which characteristics (variables) are known. The structures of the first six datasets introduced in Section 1 are given in Table 8 (overleaf).

In Section 1, we stressed one particular important feature that distinguishes between datasets. This is that variables come in different types – continuous, discrete, ordinal categorical, nominal categorical – and datasets will be made up of one or more variables of one or more types. Another feature that distinguishes between datasets is whether the observations in a dataset represent a *population* or are a *sample* from a population.

Table 8 Structure of datasets from Section 1

Example	Objects	Variables
Example 1	countries	number of operational power stations
Example 2	occupation types	number of males employed number of females employed total number of people employed
Example 3	Canadian mothers	number of children
Example 4	areas in England	percentage of sports club membership
Example 5	participants	treatment group number of training sessions gender weight change
Example 6	patients	method of tattoo removal gender size of tattoo depth of tattoo quality of removal score

Populations are not only human

The objects that form a dataset constitute a **population** if, collectively, they represent all the objects that it is possible to have. For example, the objects in Example 1 form a population – the population of countries which had at least one operational nuclear power station in 2014. This is because all such countries are listed in Table 1 (given in Example 1). Similarly, the population of a country, that is, all the people that live in a country, is indeed a population in this sense, but it is very far from being the only type of population of interest in statistics.

On the other hand, if only some objects that it is possible to have are included in the dataset, then the given objects that form a dataset are said to be a **sample** from a population. For example, the data in Example 6 can be thought of as a sample. This is because we can think of the patients included in that dataset as just some of the patients who had tattoos removed using either method 'A' or method 'B'.

Note that, with samples, it is important to think about the **underlying population**. That is, what set of objects is the set of objects we have in our dataset a subset of? For example, for the data in Example 6, the underlying population can be thought of as all patients who have tattoos removed using either method 'A' or method 'B'. Indeed, specifying the underlying population of interest can be thought of as being driven by who you wish to apply your findings to. In the tattoo removal example, this population could then even include future patients who will have their tattoos removed using method 'A' or method 'B'. Specifying an appropriate underlying population is not always easy.

Activity 4 *Sample or population?*

For each of the following datasets which you have already met in this unit, state whether you think it forms a population or a sample. Further, for those datasets which you think are samples, suggest what the underlying population is.

(a) The total number of people employed in all occupation types in the UK in the last quarter of 2015 (Example 2).

(b) The clinical trial assessing response inhibition training as a method to encourage weight loss (Example 5).

(c) The percentage of people who were members of sports clubs in different English areas (Example 4).

(d) Measurements of blood plasma β endorphin concentrations in runners who did not collapse in the Great North Run (Activity 3).

(e) The data from the Crime Survey for England and Wales 2015–16 (Exercise 2(b)).

Activity 5 *Earthquakes*

In one 24-hour period in March 2016, 21 earthquakes around the world of magnitude at least 2.5 were recorded by the US Geological Survey. For this dataset, discuss whether it forms a population or a sample and, if a sample, what the underlying population is.

As you have seen in Activities 4 and 5, for any particular dataset the definition of the underlying population is often not clear-cut. However, despite this ambiguity, in many analyses it does matter. Frequently, the goal in statistical analysis is to move beyond describing and summarising the data that have been observed to making statements about the underlying population, which has not been observed in its entirety. So then it is important to know what the underlying population actually is! Crucially, in moving from making statements about a sample to making statements about the underlying population, there is an assumption that the sample is **representative** of the underlying population. That is, properties of the underlying population are well reflected in the sample.

For some samples, the assumption of representativeness is easy to justify, for example, when a sample is taken from a population using *simple random sampling*. That is, the sample is taken by using a process by which every object in the population is given the same probability of ending up in the sample as every other object. For other samples, the assumption is more open to debate. For example, consider the question of whether the earthquakes that occur in one 24-hour period are representative of all earthquakes that occur over a year. If the 24-hour period is chosen because it was known that a particularly strong earthquake occurred during that

The USA, like the UK, has a two-house parliamentary system (called Congress in the USA). The lower house in the USA is the House of Representatives, comprising 435 elected voting members. The question is: is the US House of Representatives representative of the US population?

period, then the sample may not be representative of all the earthquakes that year. However, it is easier to argue that the sample is representative if the 24-hour period is chosen for reasons unconnected with the earthquakes that actually occurred.

Activity 6 *Representative or not?*

In Activity 4(c), it was noted that the data given in Example 4 can be regarded as information about the population of different areas in England in 2014–15. Suppose now that the following samples are taken from this population. For each sample, state whether you think the assumption of representativeness is reasonable, justifying your opinion.

(a) The areas in South East England.

(b) Six areas chosen at random from all the areas.

(c) Every sixth area when the areas are placed in alphabetical order of the names of the areas.

As you have seen, deciding whether or not a sample is representative is often a judgement call. Even selecting a sample using simple random sampling is not a guarantee that a sample will definitely be representative. It just means there is no reason to think it won't be. Ultimately, the only way to be sure that a sample is fully representative would be to gather information about the whole population. However, in most situations, this is impractical, if not impossible. So it is a judgement call that has to be made. In M248, you should assume that all samples are representative unless you are told otherwise.

Exercise on Section 2

Exercise 3 *Chondrites*

Meteorites are chunks of rock that originate from objects in space and land on the Earth's surface. Chondrites are meteorites that have not undergone particular physical processes (e.g. melting). In Section 4, data from 22 chondrites will be given. The chondrites studied were a subset of the known chondrite meteorites from around the world that were available to a particular chemical analyst in the 1950s and early 1960s. This dataset will have just one variable, the percentage of silica the chondrite contains.

(a) Why is it reasonable to regard this dataset as a sample?

(b) What is the underlying population?

(c) Is it reasonable to regard this sample as being representative?

3 Graphics

In Section 1, you began the process of getting a feel for data. You considered what type(s) of variable a dataset contains and which, if any, variables are linked. In this section, we continue this process by plotting individual variables. Such plots will be used to provide initial answers to questions such as:

- What range of values occur in the data?
- Which values are common and which are not?
- Do any data points appear particularly unusual? If so, in what way?

These questions could be answered by looking at a table of the data. However, even for small datasets such as those given in Section 1, a good plot displays the data in such a way that initial answers to these questions become available in a much more immediate way. In Subsection 3.1, you will see how *bar charts* can be used to display categorical and discrete data. Then, in Subsection 3.2, a plot useful for displaying continuous data will be discussed: the *histogram*. A third type of plot, the *boxplot*, will be introduced in Subsection 3.3; it is also primarily useful for displaying continuous data. You will get your first opportunity to use Minitab in Subsection 3.4.

3.1 Bar charts

A bar chart is a type of plot that is used to display the values of categorical and discrete variables.

In a **bar chart**, each possible category or discrete value is represented by a bar, the height of which corresponds to the number of times that category or discrete value occurs in the dataset. These numbers are often called the **frequency** of occurrence of the category or discrete value.

Example 7 *Success in removing tattoos*

In Example 6, data on the surgical removal of tattoos were listed. A bar chart of one of the variables, score, is given in Figure 1. This score is a measure of the quality of tattoo removal, measured on an ordinal scale with values '1', '2', '3' and '4'; here, '1' represents poor removal, '4' represents excellent removal, and the other two categories are intermediate.

In Figure 1, a vertical bar has been drawn of height equal to the frequency of occurrence of each score in the dataset. In fact, in this dataset, there were 14 scores of '1', 14 scores of '2', 11 scores of '3' and 16 scores of '4'. The frequencies can be read off from Figure 1 or, with more effort, worked out from Table 6. But this is not the point of the graphic. The important thing about Figure 1 is that it makes clear, at a glance, that the degrees of success of the tattoo removals are evenly spread between categories. A score of '3' occurred just slightly less often than the other scores, and a score of '4' slightly more frequently.

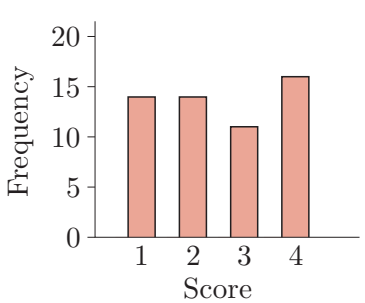


Figure 1 Success in the removal of tattoos

Example 8 *Distribution of number of children*

In Table 3 of Example 3, the number of children born to each of 35 Canadian mothers in the first half of the twentieth century were given. A bar chart of the data is given in Figure 2.

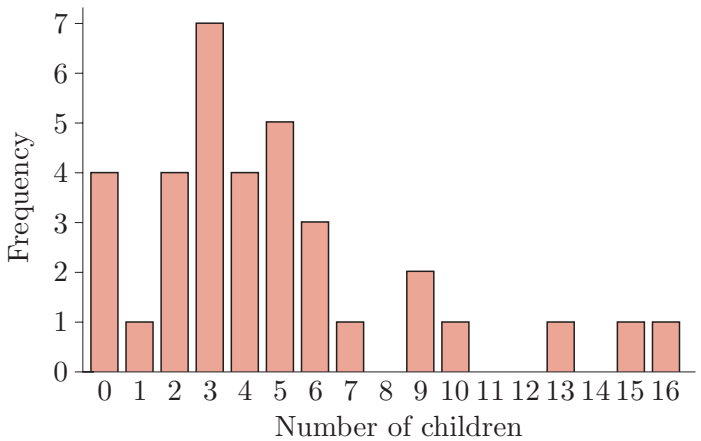


Figure 2 Numbers of children born to particular Canadian mothers

Figure 2 makes it clear that most mothers had six or fewer children but some had more (the largest number being 16). Among the smaller family sizes, most frequencies were broadly similar, with the largest frequency corresponding to 3 children. However, just one mother had a single child.

Example 7 concerned a bar chart of categorical data while Example 8 concerned a bar chart of discrete data. Notice that, in both Figures 1 and 2, there are gaps separating each of the bars. *This is an important feature of bar charts.* The gaps emphasise the fact that categories are distinct, as are discrete data values, and that there is no meaningful 'continuous' link between them.

When the categorical variable is ordinal, as was the case in Example 7, it is usual to order the bars with respect to the natural ordering. For nominal variables, there is no such natural ordering to dictate the ordering of the bars. Instead, the bars are often ordered with respect to the heights of the bars, to assist in the comparison of heights.

Activity 7 *Distribution of the UK workforce*

In Example 2, data on the number of people in employment in the UK in the last quarter of 2015 were given. A bar chart of the total numbers employed, categorised by occupation type, is shown in Figure 3. 'Occupation type' is a nominal variable, so its categories have been ordered from highest frequency on the left to lowest frequency on the right.

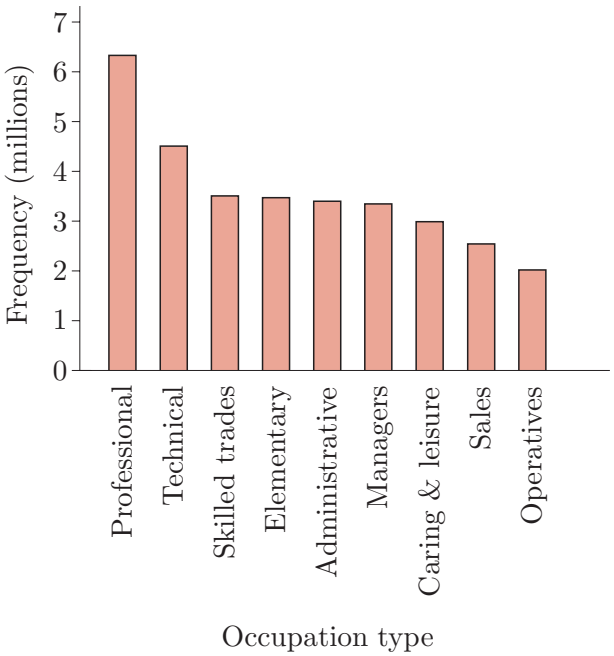


Figure 3 Employment in the UK

Using this bar chart, briefly comment on the distribution of the UK workforce at the end of 2015 across occupation types.

In each of the above bar charts, the bars are drawn vertically. This will remain the standard convention throughout the main text of this module. However, bar charts can also be displayed horizontally; the choice between vertical and horizontal bar charts is not very important and can in general be made according to convention, preference or convenience.

An issue for clothing: vertical stripes or horizontal stripes?

3.2 Frequency histograms

As you have seen in Subsection 3.1, bar charts give a visual display of the distribution of categorical and discrete variables. However, not all variables are categorical or discrete. What is to be done for continuous variables? One approach is to split the range of possible values for a continuous variable into intervals and then produce a display that is similar to a bar chart – the **histogram**. The positions at which the range is split are known as *cutpoints*; the intervals defined by the cutpoints are known as *bins* (because we're sorting the data into different 'bins' according to the data values).

Construction of frequency histograms

Let us start by considering an example.

Example 9 *Binning percentages*

In Table 4 of Example 4, the percentages of adults who were members of sports clubs in 49 English areas were given. In order to understand how a histogram is constructed, it is useful to identify an interval within which all the data lie. By scanning through Table 4, it can be seen that the percentages observed were between 16.7 (the smallest percentage observed) and 27.7 (the largest percentage observed). So one interval within which all the data lie is the interval from 16.0 to 28.0. One way of splitting up this interval is into the subintervals, or bins, 16.0–17.0, 17.0–18.0, ..., 27.0–28.0. That is, take the cutpoints to be 16.0, 17.0, 18.0, ..., 28.0. Note that the bins do not overlap, neither do they leave any gaps between them.

You might argue that, strictly speaking, the end values 16.0 and 28.0 are not really cutpoints, but it proves convenient to include them as such.

Having decided on some cutpoints, it is clear which bin a percentage belongs to as long as it is not equal to a cutpoint – which some of them are! This leaves the issue of what to do when a percentage is exactly equal to a cutpoint. For example, into which bin should the percentage 24.0 that occurs in row 3, column 1 of Table 4 be put? It turns out that it does not matter very much if such cases are put in the bin to the left of the value (for 24.0, this is bin 8, 23.0–24.0) or in the bin to the right of the value (for 24.0, this is bin 9, 24.0–25.0); what matters is that whatever is done is done consistently for all cases which are equal to cutpoints. For example, if the percentage 24.0 is put in bin 9 (to its right), then 25.0 (row 5, column 3 of Table 4) should be put into bin 10 (to its right), and the two further cases of 24.0 (row 5, column 2 and row 1, column 6 of Table 4) should also be put in bin 9.

This is the convention that we have chosen to use throughout this module: *if a data point has a value equal to a cutpoint, put it in the bin also containing higher values, to the right of the cutpoint.*

Table 9 Frequencies of observed percentages in the sports club membership data

Bin	Values	Frequency
1	16.0–17.0	2
2	17.0–18.0	2
3	18.0–19.0	2
4	19.0–20.0	7
5	20.0–21.0	6
6	21.0–22.0	7
7	22.0–23.0	10
8	23.0–24.0	4
9	24.0–25.0	5
10	25.0–26.0	2
11	26.0–27.0	0
12	27.0–28.0	2

Following the convention and then counting the number of observed percentages that fall into each of the bins, we get Table 9.

The histogram corresponding to the bins and their frequencies given in Table 9 is shown in Figure 4. As in a bar chart, each possible bin – the continuous data analogue of a categorical group – is represented by a bar, the height of which corresponds to the frequency with which values in that bin occur in the dataset. It shows that most areas have sports club memberships of between about 19% and 25% of adults, with some having smaller percentages and a few higher. The bin associated with the greatest number of areas has between 22% and 23% membership.

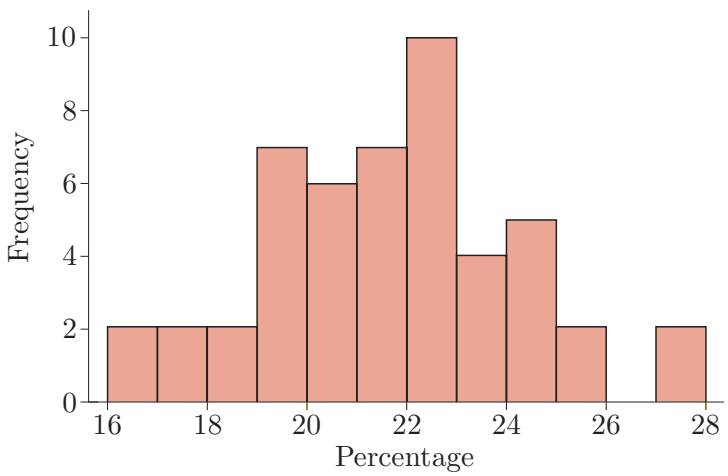
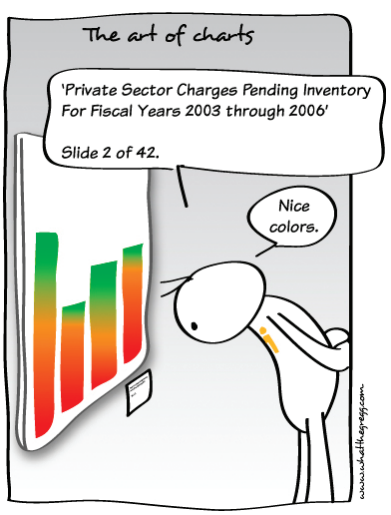


Figure 4 A histogram of the sports club membership data

As you have seen in Example 9, having grouped a set of continuous data into bins, a plot known as a histogram can be drawn where the number of cases in each bin is represented by a bar. Thus bar charts and histograms are similar. However, there are two key differences between bar charts and histograms.

1. On bar charts, gaps are left between bars. On histograms, there are no such gaps. This is because, with continuous data, all values in the range of interest are possible.
2. On bar charts, the *heights* of the bars correspond to the numbers of cases in the groups. On histograms it is the *areas* of the bars that correspond to the numbers of cases in the bins.

That said, often the cutpoints are defined so that *the widths of all the bins are the same*. This is another convention that we will follow throughout this module. For example, all the bins given in Table 9 are of the same width, namely, one percentage point. When bins are all the same width, the heights of the bars are also proportional to the numbers of cases in the bins. And so we can, for now, forget about areas of bars and simply make the height of each bar *equal* to the frequency associated with the bin (which is just like a bar chart, after all!). Such histograms are known as **frequency histograms**.



Area = height × width

You will meet another sort of histogram, a unit-area histogram, in Subsection 5.2.

Before considering the interpretation of frequency histograms, we will have to consider further complications that arise in their construction. You will probably have realised that there is more than one way of splitting the range of possible values of continuous data into a set of bins using a series of cutpoints. There is choice about the position of the starting point (so long as it is not greater than the smallest observed value) and of the distance between cutpoints (the widths of the bins). These choices can make quite a lot of difference to the resulting frequency histogram.

Activity 8 *Other binnings of percentages*

You could use the frequencies given in Table 9 rather than going back to the data in Table 4.

For each of the following sets of cutpoints, give the corresponding table of frequencies for the sports clubs membership data. (Where an observed percentage is the same as a cutpoint, place the percentage in the group containing higher values, to its right.)

(a) Cutpoints: 16.0, 18.0, 20.0, 22.0, 24.0, 26.0, 28.0.

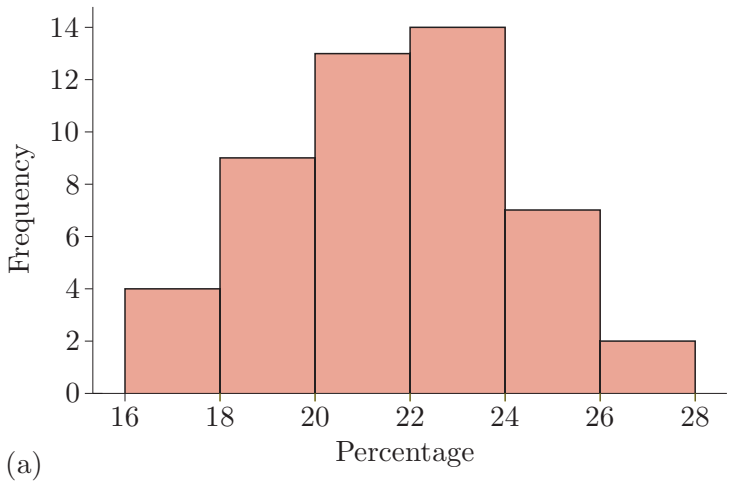
(b) Cutpoints: 16.0, 20.0, 24.0, 28.0.

(c) Cutpoints: 15.0, 17.0, 19.0, 21.0, 23.0, 25.0, 27.0, 29.0.

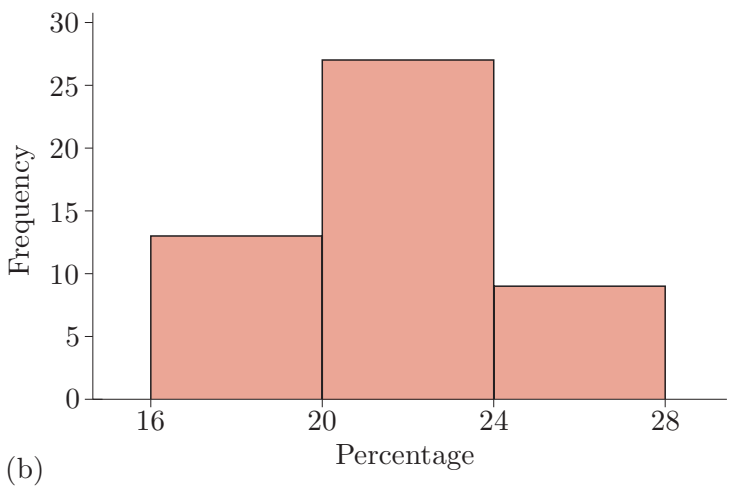
(Histograms using these cutpoints will be considered in Activity 9.)

Example 10 *Frequency histograms of the sports club membership data*

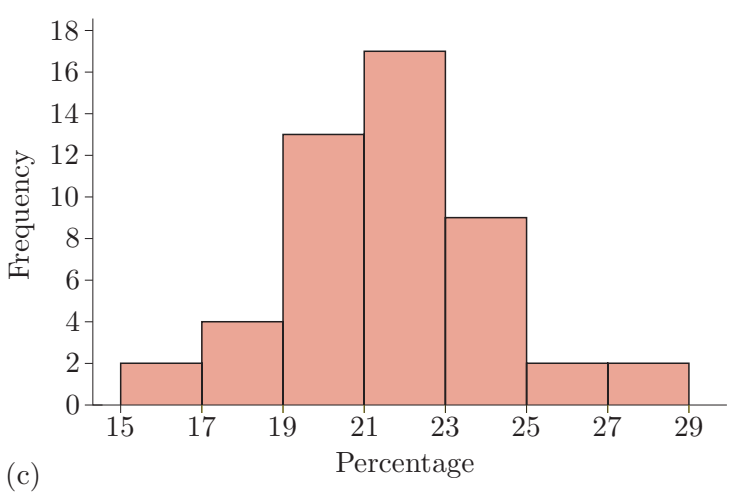
Figure 4 in Example 9 is a frequency histogram of the sports club membership data given in Table 4, using the binning scheme given in Table 9. Figure 5 shows three more frequency histograms of the sports club membership data. Each of these is based on a different binning scheme: Figure 5(a) on the bins given in Activity 8(a); Figure 5(b) on the bins given in Activity 8(b); and Figure 5(c) on the bins given in Activity 8(c).



(a)



(b)



(c)

Figure 5 More histograms of the sports club membership data

Activity 9 *Which frequency histogram?*

Figures 4, 5(a), 5(b) and 5(c) display four frequency histograms of the sports club membership data, and Figure 6 below a fifth. (Figure 6 corresponds to a frequency histogram of these data automatically produced by Minitab.)

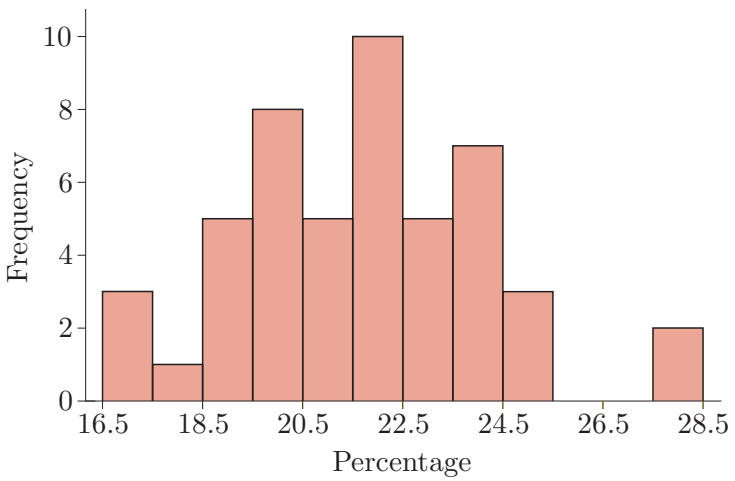


Figure 6 A fifth histogram of the sports club membership data

(a) Which of Figures 4, 5(a), 5(b), 5(c) and 6 are based on the same starting position as one another?

(b) By comparing the histograms with the same starting position that you identified in part (a), what kind of effect do you think the choice of bin width has on the appearance of a frequency histogram?

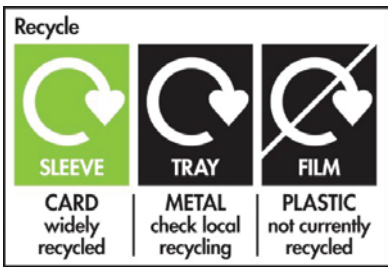
(c) Which of Figures 4, 5(a), 5(b), 5(c) and 6 are based on the same bin width as one another?

(d) By comparing the histograms with the same bin width that you identified in part (c), what kind of effect do you think the choice of starting position has on the appearance of a frequency histogram?

As suggested in the solution to Activity 9, the choice of cutpoints, or equivalently of the starting position and width of bins, in a histogram can have a substantial effect on the appearance of a histogram and hence its interpretation. In particular, the smaller the width of the bins, the more detail is retained in the histogram. Too much detail, however, can obscure the main messages of a histogram, so some intermediate value of bin width is usually chosen which attempts to balance level of detail on the one hand with the main features of the distribution of the data on the other.

The effect of changing starting position is less clear-cut but can also be substantial.

We will pursue this issue no further in this module except to emphasise that when you are provided with a histogram of some data, particular choices have been made – hopefully in a sensible way, either automatically by a computer program such as Minitab or by reasonable choices made by a person. Other choices (of starting position and/or bin width) may well have given rise to a histogram with a different appearance. It therefore really makes sense only to talk of '*a* histogram' of some data rather than '*the* histogram' of the data. It also means that one should not over-interpret every little bump and dip in a histogram, a point taken up again in the next passage.



Guidance on what goes in what bin?

Interpretation of frequency histograms

Looking at a frequency histogram gives an insight into the shape of the distribution of the data. When interpreting a histogram, it is usual to consider the following two points, in particular:

1. the number of modes
2. whether the distribution of the data is symmetric and, if not, whether the data are left-skew or right-skew.

Statisticians often just say 'the data are' rather than 'the distribution of the data is'.

These two aspects of histogram shape are explained next.

A **mode** in a histogram corresponds to a peak in the heights of the bars. Note that a mode does not just refer to the tallest bar or bars. It is possible for a bar to correspond to a mode without it being the tallest bar overall: a bar corresponding to a mode just needs to be taller than the bars either side of it. The data are **unimodal** if there is just one mode, **bimodal** if there are two modes, and **multimodal** if there are more than two modes.

If you are familiar with the terminology of mathematical optimisation, you will see that modes include both local and global maxima in a histogram.

A histogram is **symmetric** if the pattern in the heights of the bars appears to be symmetric around some central point. Asymmetry (non-symmetry) is often called **skew** in statistics. Left-skew and right-skew refer to the manner in which the data decline away from the central point: if the heights of the bars decline away to zero more slowly on the left-hand side compared with the right-hand side, the data are said to be **left-skew**. Conversely, if the heights of the bars decline away to zero more slowly on the right-hand side compared with the left-hand side, the data are said to be **right-skew**. The direction of skew is of interest mainly when the data are both unimodal and asymmetric.

'Skew' is often written in other texts as 'skewed'.

A pictorial summary of these two features is given in the following figures.

Histograms of different distribution types

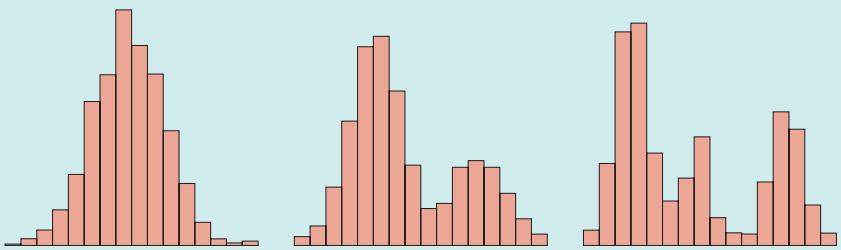


Figure 7 Modality of histograms

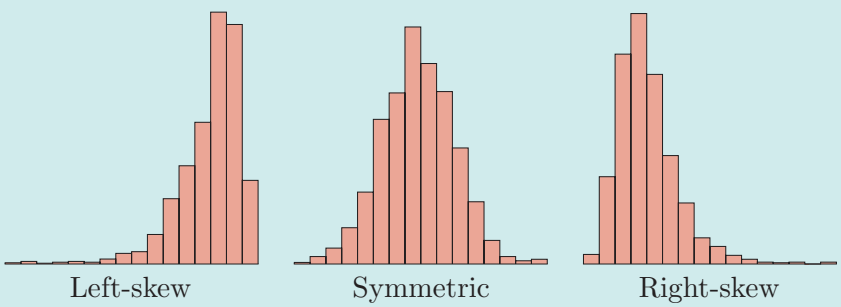


Figure 8 Symmetry and skew of histograms

Notice that the 'symmetric' histogram in Figure 8 is not exactly symmetric. (For example, the bar immediately to the right of the highest bar is higher than the bar directly to the left of the highest bar.) Approximate symmetry is all we require to declare a histogram symmetric. Similarly, small modes in a histogram are sometimes discounted. (See Example 11 below.) This kind of approximation is justified at least partly because detailed characteristics of a histogram might be changed if its starting position and/or bin width were to be changed.

Example 11 *Interpreting a histogram*

In Figure 4 of Example 9, a histogram of the sports club membership data was given.

In this histogram, strictly speaking, there are three modes, at the bins centred at 19.5, 22.5 and 24.5, indicating that the data are multimodal. However, only the middle peak, at roughly 22.5, clearly stands out, so it could be argued that the data are unimodal.

Arguably, the bars on the left-hand side of the histogram fall away more slowly than those on the right-side. So there is some suggestion that the data are left-skew. But the effect is slight and one could just as well claim that the histogram is, approximately, symmetric.

Example 12 *Interpreting another histogram*

The histogram in Figure 9 was provided on behalf of the Intergovernmental Panel on Climate Change in 2000. It summarises the predictions of global primary energy consumption in the year 2100 under $n = 127$ different scenarios. The units, EJ, are exajoules, that is, 10^{18} joules.

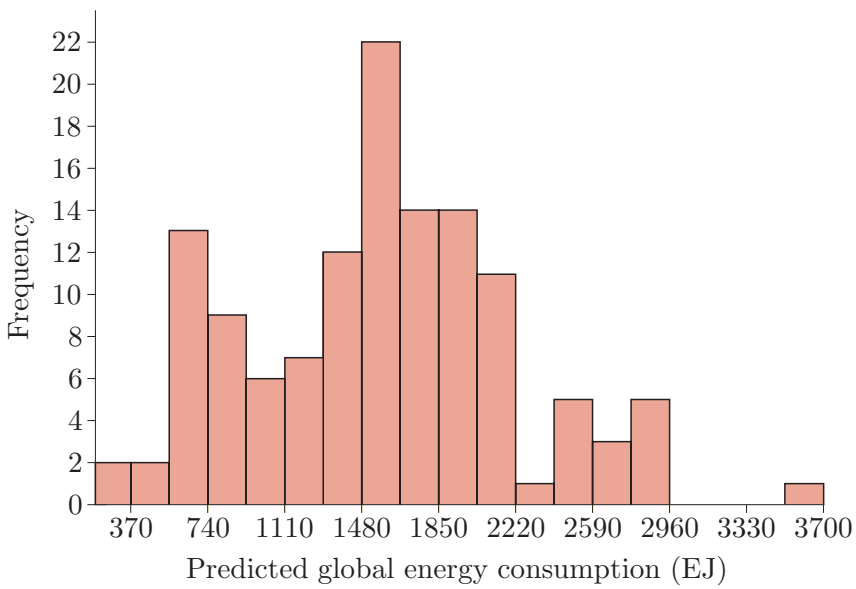


Figure 9 Predicted global energy consumption in 2100 (EJ)

(Source: Nakicenovic, N. and Swart, R. (eds) (2000) *Emissions Scenarios*, Cambridge, Cambridge University Press)

The histogram shows a multimodal distribution. Strictly speaking, there are five modes; most statisticians would probably settle on three modes and an outlier! There is a first mode around 700 EJ or so, reflecting a number of scenarios resulting in relatively low energy consumption predictions. The largest mode is at around 1500 EJ, with many scenarios resulting in predicted energy consumption values between about 1300 EJ and 2200 EJ. The third and fourth modes (counting from the left) can be argued to be part of a cluster of predictions of high energy consumption values (between about 2400 EJ and 2900 EJ). And finally a single scenario gives rise to the outlyingly high prediction (between about 3500 EJ and 3700 EJ).

All but two scenarios result in energy consumptions greater than the 1991 reference value of 370 EJ.

Questions of skew are not really relevant, given the multimodality of these data.

Activity 10 *Describing histograms*

Briefly describe the shapes of the following two histograms.

The highest life expectancies, of 87 years, are in Hong Kong and Japan

(a) A histogram of the life expectancy at birth for girls born in 2013 in various countries across the world is shown in Figure 10.

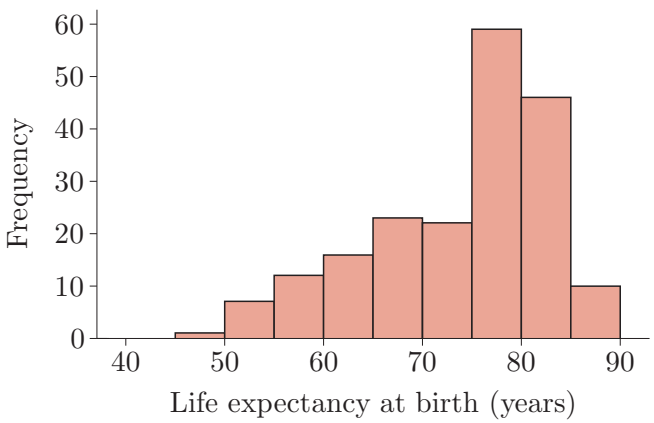


Figure 10 Life expectancy at birth (years)

(Source: World Bank, http://data.worldbank.org/indicator/SP.DYN.LE00.FE.IN)

Old Faithful: little to see between eruptions!

(b) The waiting times between almost 300 eruptions of the Old Faithful geyser in Yellowstone National Park, USA, were recorded in August 1978. A histogram of these data is given in Figure 11.

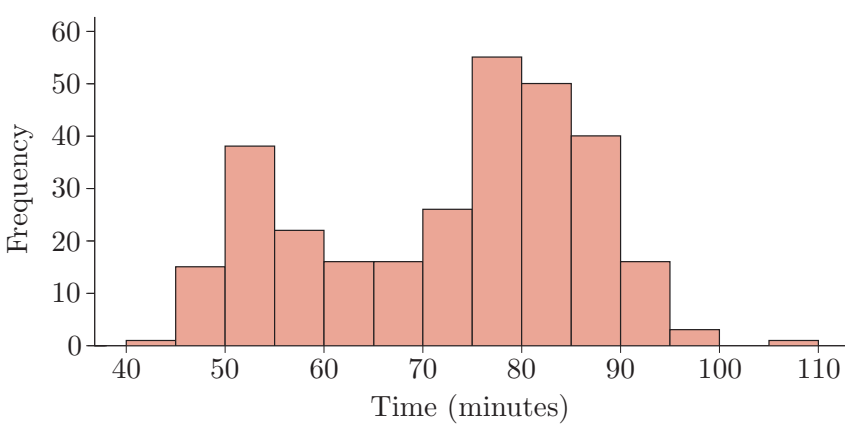


Figure 11 Waiting times between eruptions (minutes)

(Source: Azzalini, A. and Bowman, A.W. (1990) 'A look at some data on the Old Faithful geyser', *Applied Statistics*, vol. 39, no. 3, pp. 357–65)

3.3 Boxplots

Histograms are not the only way of displaying a continuous variable. Another type of plot that is frequently used is the **boxplot** (also known as a box-and-whisker plot). Boxplots consist of a number of elements.

These elements utilise some numerical summaries of the data, specifically their median and quartiles, the details of which need not concern you now but will be reviewed in Section 4.

- A 'box' – the length of which indicates the range of values over which the middle 50% of the data lie. That is, the box ranges from the lower quartile of the data to their upper quartile.
- 'Whiskers' – lines extending out from the box indicating the range of values for the rest of the data (except potential outliers – observations that do not appear to be following the same pattern as the rest of the data).
- Individual points – points beyond the whiskers which are sufficiently different to the rest of the data that they can be regarded as potential outliers.
- A line in the box – indicating the value of the middle (sample median) of the data.

Boxplots

The general structure of a boxplot is shown in Figure 12.

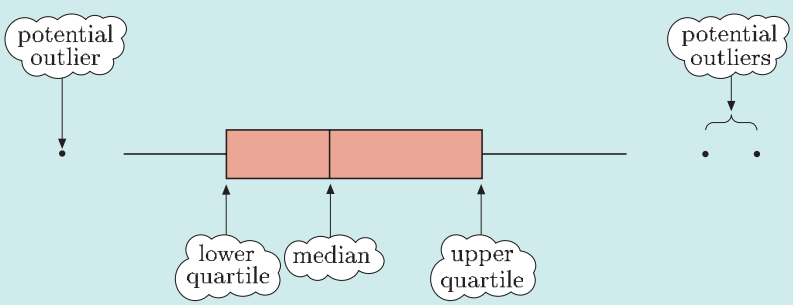


Figure 12 Schematic of a boxplot

Example 13 *Boxplot of sports club membership*

Figure 13 shows a boxplot of the sports club membership data given in Table 4 (Example 4).

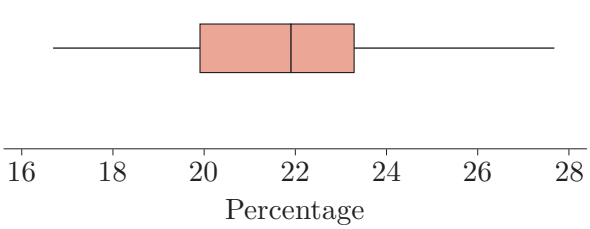
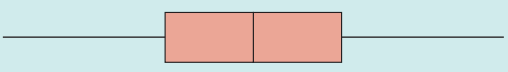


Figure 13 Boxplot of the sports club membership data

On this boxplot, no individual points are shown. So there are no areas that are to be regarded as potential outliers. The middle 50% of membership percentages lie between about 20 and 23; those values are approximately the limits of the central box. Beyond that, all the membership percentages lie between about 17 and 27.5; those values are approximately the limits of the whiskers.

From boxplots it is not possible to tell how many modes the distribution of a variable might have. However, they do provide information about the symmetry or otherwise of the data. This is indicated by the relative lengths of the whiskers and the relative lengths of the box 'halves' either side of the median line. The following box provides examples.

Boxplots and the shape of data



Symmetric

Left-skew

Right-skew

Asymmetric (but not clearly left- or right-skew)

Asymmetric (but not clearly left- or right-skew)

Figure 14 Symmetry and skew of boxplots

The second boxplot in Figure 14 reflects left-skew because both the left box half is longer than the right box half and the left whisker is longer than the right whisker. The third boxplot indicates right-skew in a vice versa way. In contrast, the fourth and fifth boxplots in Figure 14 both have their boxes and their whiskers suggesting opposing types of skewness and hence indicate a complicated form of skew, or more general structure, in the data. The main message of plots like these last two is that while the data are not symmetric, neither can they be described simply as left-skew or right-skew.

Example 14 *Shape of the sports clubs membership data*

You might argue that in Figure 13, the right whisker is longer than the left whisker and the left box half is longer than the right box half. So from the boxplot the distribution of percentages of adults who were members of sports clubs appears to be asymmetric, but not clearly left- or right-skew. This is slightly different to the conclusion reached in Example 11. However, such ambiguity is quite reasonable given that any lack of symmetry displayed by these data is not at all pronounced.

Activity 11 *Beta endorphin levels in collapsed runners*

In Activity 3, some data on the β endorphin levels of runners in the Great North Run was described. A boxplot of the data for the collapsed runners is given in Figure 15.

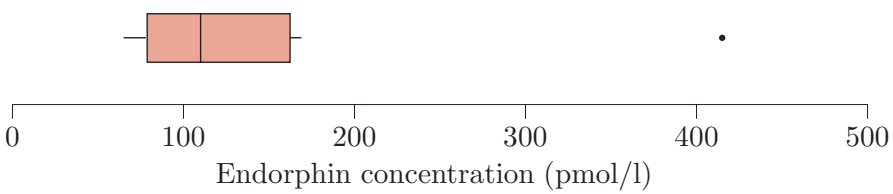


Figure 15 β endorphin levels in collapsed runners

Use this boxplot to describe the shape of the distribution of these data.

In the 1954 Empire Games marathon in Vancouver, British runner Jim Peters led the field by 17 minutes but collapsed repeatedly within about 200 metres of the finish and could not complete the race

3.4 Introducing Minitab

It is now time to transfer your attention to Computer Book A and work through Chapters 1 and 2 of it. In Chapter 1, you will be introduced to the data analysis software Minitab. Among other things, you will learn how to produce bar charts using Minitab. Then, in Chapter 2, you will use Minitab to produce frequency histograms and boxplots of datasets.

Refer to Chapters 1 and 2 of Computer Book A for the rest of the work in this section.

Exercises on Section 3

Exercise 4 *Interpreting a bar chart*

Exercise 1 described some variables collected in the Crime Survey for England and Wales (CSEW). One of these was the length of time the respondent has been living in their current area (in years), grouped as '< 1', '1–2', '2–3', '3–5', '5–10', '10–20' or '20+'. A bar chart of the data on this variable that were collected in 2007–08 (when the CSEW was called the British Crime Survey) is displayed in Figure 16.

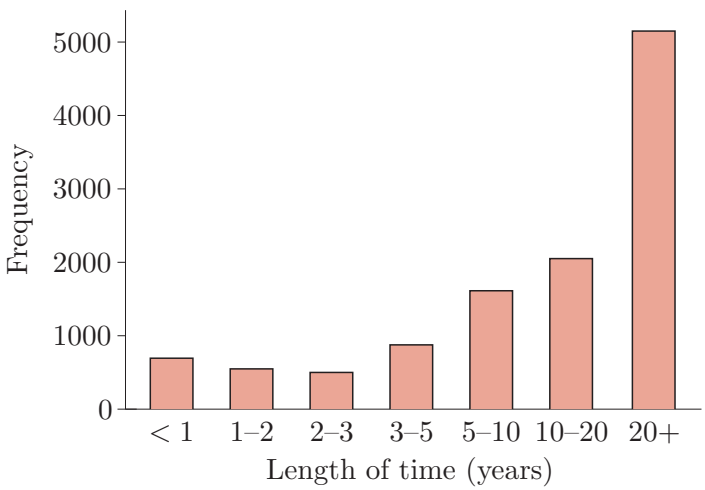


Figure 16 Lengths of time lived at current address

Using this bar chart, comment on the distribution of the lengths of time respondents had been living in their current area.

Exercise 5 *Interpreting a histogram*

From the variables collected in the British Crime Survey 2007–08, researchers derived a new variable, the 'level of worry about being a victim of personal crime'. The way this variable was constructed means that it is effectively continuous. The scale was set so that a value of 0 was a 'middling' value, and the higher the value, the more worry the respondent had. A histogram of the values of this variable is given in Figure 17.

Using this histogram, comment on the distribution of the levels of worry reported.

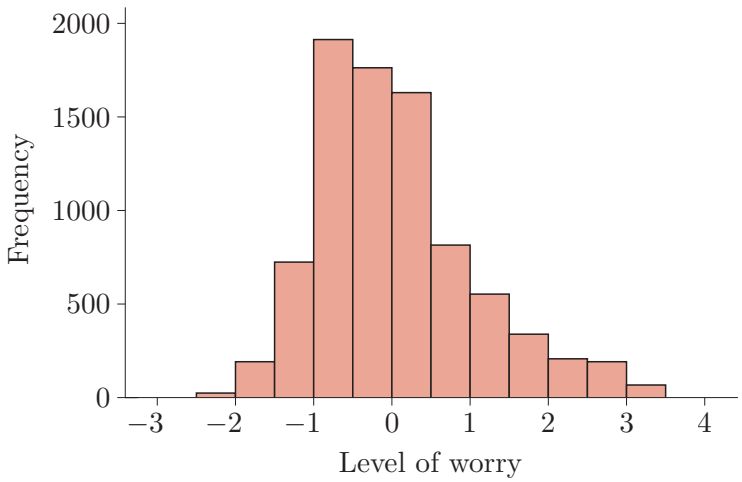


Figure 17 Level of worry about being a victim of personal crime

Exercise 6 *Interpreting a boxplot*

This exercise concerns the data briefly described and considered in Activity 5. These consist of the magnitudes of 21 earthquakes of magnitude at least 2.5 recorded in a 24-hour period in March 2016. Figure 18 is a boxplot of the magnitudes of these earthquakes.

Using this boxplot, describe the distribution of the magnitudes of these earthquakes.

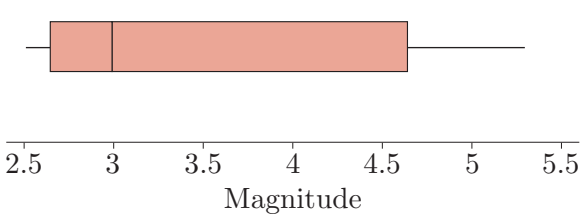


Figure 18 Magnitudes of earthquakes occurring in one 24-hour period

Taking protective action in an earthquake

4 Numerical summaries

In Section 3, you have seen how plots can be used to get an initial feel for the data. Using those plots it is possible, among other things, to get some idea of the 'typical' or 'average' value, and of the 'dispersion' or 'variability' of the values. In this section, we will review methods for adding numerical values to these statements, that is, for calculating useful **numerical summaries**.

Measures of *location* – a term covering typical or average values – are discussed in Subsection 4.1, and measures of *spread* – a term covering dispersion or variability of the values – in Subsection 4.3. Spanning the two is a subsection on sample quartiles (Subsection 4.2). In Subsection 4.5, you will use Minitab to calculate the numerical summaries described in this section.

Throughout this section, we will concentrate on summarising a single variable in a dataset, and we will assume that this variable is either continuous or discrete (that is, it is numerical rather than categorical). If more than one continuous or discrete variable is to be summarised numerically, then this can be done, at least partially, by applying the techniques introduced in this section to each variable in turn.

A first, unambiguous, numerical summary of a dataset is the number of observations it contains. This quantity is called the **sample size** and is very often denoted by n.

4.1 Measures of location

Measures of location give ways of calculating a value that is 'typical' of, or 'centrally located' in, a dataset. In this subsection, we will consider two such measures: the sample mean and the sample median. Here the prefix 'sample' is important. It signifies that we are calculating these quantities for the sample of observations that we actually have at hand. In later units, you will meet 'population' versions of these quantities, that is, quantities that, theoretically at least, are calculated for the underlying population.

The sample mean

As you are probably already aware, the sample mean is the sum (or total) of the data values divided by the number of observations (the sample size). The sample mean is, therefore, probably what most people think of when they think of the 'average' of the data values. More formally, this is written in the following way.

This mean or average is also called the arithmetic mean or arithmetic average.

The sample mean

If the n values in a dataset are denoted $x_1, x_2, \ldots, x_n$, then the **sample mean**, which is denoted $\overline{x}$, is given by

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The symbol $\overline{x}$ denoting the sample mean is read 'x-bar'.

Σ is the Greek upper-case letter sigma.

Recall that the symbol Σ is used to mean 'the sum of'. The expression $\sum_{i=1}^{n}$ is read 'sigma i equals 1 to n', and $\sum_{i=1}^{n} x_i$ means the sum of the terms $x_1, x_2, \ldots, x_n$.

Example 15 *Calculating a sample mean*

In Table 4, the percentages of adults who were members of sports clubs in 49 areas covering the whole of England were given. From this population, a random sample of six areas has now been selected. The percentage of adults who were members of sports clubs for the selected areas is as follows.

19.1 17.4 23.7 22.3 16.7 22.6

If $x_1, x_2, \ldots, x_6$ represents all the cases in this sample, then just working along the list, $x_1 = 19.1$, $x_2 = 17.4$, $x_3 = 23.7$, $x_4 = 22.3$, $x_5 = 16.7$ and $x_6 = 22.6$.

It doesn't actually matter which value corresponds to which x.

The mean of these percentages is therefore

$$\overline{x} = \frac{1}{6}\sum_{i=1}^{6} x_i = \frac{1}{6}(x_1 + x_2 + x_3 + x_4 + x_5 + x_6)$$
$$= \frac{1}{6}(19.1 + 17.4 + 23.7 + 22.3 + 16.7 + 22.6) = \frac{121.8}{6} = 20.3.$$

Thus the sample mean percentage is 20.3%.

Activity 12 *Calculating another sample mean*

In Activity 3, the β endorphin concentrations of 11 runners who collapsed during the Great North Run were given. For convenience, the data are given again below.

66 72 79 84 102 110 123 144 162 169 414

Calculate the sample mean.

The sample median

Calculation of the sample median starts out by placing all the observations in order. In M248, we will assume that whenever a variable is ordered, the values will be placed in *increasing* order. The **sample median** is defined as 'the middle value' in the dataset, a simple definition that, as you probably already know, is not always quite as simple as it seems. The idea is that the sample median is a value for which approximately the same number of the data values are smaller than it as are larger than it.

Before writing the definition of the sample median, it is first necessary to distinguish between the order that the observations arrived, or were given, in and their order after they are placed in ascending order. Suppose that, for a sample of n cases, the values of a variable are $x_1, x_2, \ldots, x_n$. We then denote by $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ the same values when placed in ascending order. So $x_{(1)}$ is the smallest value in the dataset, $x_{(2)}$ is the second smallest value and so on; in particular, $x_{(n)}$ – the nth smallest value – is the largest value in the dataset.

In some texts, you will see ordered data represented as $x_{1;n}, x_{2;n}, \ldots, x_{n;n}$.

Activity 13 *Ordering observations*

In Example 15, the percentages of adults who were members of sports clubs in a sample of six areas were given as follows:

$$x_1 = 19.1, \quad x_2 = 17.4, \quad x_3 = 23.7,$$
$$x_4 = 22.3, \quad x_5 = 16.7, \quad x_6 = 22.6.$$

(a) Reorder these values so that they are in ascending order.

(b) Which observations, out of $x_1, x_2, \ldots, x_6$, correspond to each of the following?

(i) $x_{(1)}$

(ii) $x_{(6)}$

(iii) $x_{(4)}$

The sample median is defined as follows.

The sample median

Let a dataset $x_1, x_2, \ldots, x_n$ be reordered as $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. Then the **sample median**, m, is given by

$$m = x_{\left(\frac{1}{2}(n+1)\right)}.$$

If the sample size n is odd, then $n+1$ is even and the number $\frac{1}{2}(n+1)$ is an integer. For instance, if $n = 5$, then $\frac{1}{2}(n+1) = \frac{6}{2} = 3$, while if $n = 27$, then $\frac{1}{2}(n+1) = \frac{28}{2} = 14$.

The central reservation splits a divided highway into two equal halves and is called the median in the USA and Australia

If the sample size n is even, then $n+1$ is odd and the number $\frac{1}{2}(n+1)$ is not an integer but has a fractional part equal to $\frac{1}{2}$. For instance, if $n = 28$, then $\frac{1}{2}(n+1) = \frac{29}{2} = 14\frac{1}{2}$. The sample median is defined to be $x_{\left(\frac{1}{2}(n+1)\right)} = x_{\left(14\frac{1}{2}\right)}$, but no member of the ordered dataset has label '$\left(14\frac{1}{2}\right)$', since all are labelled by integers.

The way round this problem is to interpret $x_{\left(14\frac{1}{2}\right)}$ as 'the number halfway between $x_{(14)}$ and $x_{(15)}$', which is in fact the average of $x_{(14)}$ and $x_{(15)}$; that is,

$$x_{\left(14\frac{1}{2}\right)} = x_{(14)} + \frac{1}{2}\left(x_{(15)} - x_{(14)}\right) = \frac{1}{2}\left(x_{(14)} + x_{(15)}\right).$$

In general, since $\frac{1}{2}(n+1) = \frac{n}{2} + \frac{1}{2}$, for n even the median is halfway between $x_{\left(\frac{n}{2}\right)}$ and $x_{\left(\frac{n}{2}+1\right)}$, that is, the median is

$$m = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right). \tag{1}$$

This is all much simpler when dealing with numbers than the above explanation in symbols might suggest!

Example 16 *Calculating a sample median*

In Example 15, the mean percentage sports club membership for a sample of areas was calculated. Here, we will calculate the sample median of the same dataset.

For this sample, $n = 6$. So $\frac{1}{2}(n+1) = \frac{7}{2} = 3\frac{1}{2}$, and the median is $m = x_{(3\frac{1}{2})}$, the value halfway between $x_{(3)}$ and $x_{(4)}$. So, using the ordered data values given in the solution to Activity 13,

$$m = \frac{1}{2}\left(x_{(3)} + x_{(4)}\right) = \frac{1}{2}(19.1 + 22.3) = \frac{41.4}{2} = 20.7.$$

Thus the sample median percentage is 20.7%.

Activity 14 *Calculating more sample medians*

(a) Calculate the median β endorphin concentration for 11 runners who collapsed during the Great North Run. (These data were recently repeated in Activity 12.)

(b) Table 10 gives the percentages of silica found in each of 22 chondrites. (Chondrites are a type of meteorite, as explained in Exercise 3.)

Table 10 Silica content of chondrites

20.77	22.56	22.71	22.99	26.39	27.08	27.32	27.33
27.57	27.81	28.69	29.36	30.25	31.89	32.88	33.23
33.28	33.40	33.52	33.83	33.95	34.82		

(Source: Ahrens, L.H. (1965) 'Observations on the Fe-Si-Mg relationship in chondrites', *Geochimica et Cosmochimica Acta*, vol. 29, no. 7, pp. 801–6)

Calculate the median percentage of silica in these chondrites.

Mean or median?

In this subsection, we have reviewed two measures of location: the (sample) mean and the (sample) median. This then raises the question of which one to use.

In many situations, it does not make much difference which is chosen. However, if the observed data include some extreme values, or even are just very skew, then the value of the median is likely to be more 'typical' of the majority of the data than is the value of the mean. This is because the median is a **resistant** measure of location, whereas the mean is not. That is, the value of the mean can be heavily influenced by one or two extreme values, whereas the value of the median is not.

This property is also known as *robustness*.

Activity 15 *Typical β endorphin concentration for collapsed runners*

In Activities 12 and 14(a), you calculated the mean and median β endorphin concentrations for 11 runners who collapsed during the Great North Run. In Activity 11, it was observed that the highest observation could be regarded as an outlier. So in this activity you will investigate what difference it makes to measures of location if this observation is dropped. (The data are in both Activities 3 and 12.)

(a) Calculate the mean β endorphin concentration for the 10 observations that are not regarded as outliers.

(b) Calculate the median β endorphin concentration for the 10 observations that are not regarded as outliers.

(c) Compare your answers to parts (a) and (b) to the mean and median based on all the observations (which are 138.6 pmol/l and 110 pmol/l, respectively). Does this support the claim that the median is a resistant measure of location, but the mean is not?

Sample size also plays a role when considering the robustness or otherwise of sample measures of location (or other summary measures obtained from data): generally speaking, the smaller the sample size, the more influence an outlier will have on a sample measure.

Resistance to outliers is not the only consideration when comparing the sample mean with the sample median; some other considerations (which we will not go into here) favour the mean over the median. It should also be said that it does no harm, especially when using a computer, to calculate both measures of location; if the sample mean and sample median are similar, then you have an especially good idea of the 'typical' value in the data; if the sample mean and sample median are considerably different, then you have an indication that there is something in the data – perhaps outlier(s), perhaps skewness – that is leading to such a difference.

4.2 Sample quartiles

As you have seen, the sample median is essentially the middle value after having placed the data values in order. Other positions in this ordered list are also useful to consider. In particular, two such positions are the values one-quarter of the way along the list and three-quarters of the way along; these are the **sample lower (first) quartile** and the **sample upper (third) quartile**. The idea here is that the sample lower quartile has approximately three times as many data values larger than it than are smaller than it. Conversely, the sample upper quartile is the value which has approximately three times as many data values smaller than it than are larger than it.

When there are n observations, $n = 3, 4, \ldots$, these sample quartiles are defined to correspond to the values at positions $\frac{1}{4}(n+1)$ and $\frac{3}{4}(n+1)$ along the ordered list.

Sample quartiles are not defined if $n = 1$ or 2.

The sample quartiles

Let a dataset $x_1, x_2, \ldots, x_n$, $n = 3, 4, \ldots$, be reordered as $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. Then the **sample lower quartile**, q_L, is given by

$$q_L = x_{\left(\frac{1}{4}(n+1)\right)},$$

and the **sample upper quartile**, q_U, is given by

$$q_U = x_{\left(\frac{3}{4}(n+1)\right)}.$$

The numbers $\frac{1}{4}(n+1)$ and $\frac{3}{4}(n+1)$ are both integers only when $n+1$ is divisible by 4. In some cases, these numbers are, again, of the form 'something and a half', which we already know how to define in the manner of the sample median when n is even. Denoting the 'something' by k, we set

$$x_{\left(k+\frac{1}{2}\right)} = x_{(k)} + \frac{1}{2}\left(x_{(k+1)} - x_{(k)}\right) = \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right).$$

In all other cases, the definitions lead to the problem of how to define the value that is 'something and a quarter' and 'something and three-quarters' of the way along the ordered list. There are different definitions that can be (and are) used. The definitions that will be used in M248 are the following:

These are also the definitions used by Minitab.

$$x_{\left(k+\frac{1}{4}\right)} = x_{(k)} + \frac{1}{4}\left(x_{(k+1)} - x_{(k)}\right),$$

$$x_{\left(k+\frac{3}{4}\right)} = x_{(k)} + \frac{3}{4}\left(x_{(k+1)} - x_{(k)}\right).$$

In words, $x_{\left(k+\frac{1}{4}\right)}$ is taken to be the value that is one-quarter of the way between $x_{(k)}$ and $x_{(k+1)}$. Similarly, $x_{\left(k+\frac{3}{4}\right)}$ is taken to be the value that is three-quarters of the way between $x_{(k)}$ and $x_{(k+1)}$.

Example 17 *Calculating sample quartiles*

In Example 16, the median percentage of people who were members of sports clubs was calculated for a sample of six areas in England, using data given in Example 15; it turned out to be 20.7%. As has already been noted, for these data $n = 6$. So

$$q_L = x_{\left(\frac{1}{4}(n+1)\right)} = x_{\left(\frac{7}{4}\right)} = x_{\left(1\frac{3}{4}\right)}$$

and

$$q_U = x_{\left(\frac{3}{4}(n+1)\right)} = x_{\left(\frac{21}{4}\right)} = x_{\left(5\frac{1}{4}\right)}.$$

This means that

$$q_L = x_{(1+\frac{3}{4})} = x_{(1)} + \frac{3}{4}\left(x_{(2)} - x_{(1)}\right)$$
$$= 16.7 + \frac{3}{4}(17.4 - 16.7) = 16.7 + 0.525 = 17.225$$

and

$$q_U = x_{(5+\frac{1}{4})} = x_{(5)} + \frac{1}{4}\left(x_{(6)} - x_{(5)}\right)$$
$$= 22.6 + \frac{1}{4}(23.7 - 22.6) = 22.6 + 0.275 = 22.875.$$

So the lower quartile is approximately 17.2%, and the upper quartile is approximately 22.9%.

Activity 16 *Sample quartiles of the β endorphin concentrations for collapsed runners*

In Activity 15, you compared the mean and median β endorphin concentrations for collapsed runners, including and excluding the outlier. In this activity, you will calculate the quartiles for the same data, again including and excluding the outlier. For convenience, the data are given once again, in ordered form, below.

66 72 79 84 102 110 123 144 162 169 414

(a) Calculate the quartiles based on all 11 observations.

(b) Calculate the quartiles if the outlier, 414 pmol/l, is excluded from the calculation.

(c) Compare your answers to parts (a) and (b). Does it matter if the outlier is included in the calculations?

If the central reservation is the median, are the lane dividers in a dual carriageway the quartiles?

You should bear in mind that, by definition, it must always be the case that

$$q_L \leq m \leq q_U.$$

If your calculations do not satisfy these relationships, something is wrong with your calculations.

Activity 17 *Sample quartiles and the boxplot*

In a boxplot, the limits of the central box correspond to the sample quartiles. Using the basic idea of the sample quartiles, justify the statement that 'the length of the box indicates where the middle 50% (approximately) of the data lie'.

4.3 Measures of spread

Subsection 4.1 was concerned with measures of location. This subsection is concerned not with summarising what value an observation might typically take, but instead with how spread out the observations are. That is, we wish to obtain measures that take larger values whenever the data are more spread out and smaller values whenever they are less spread out. We consider two of the prime ways of measuring what we need to measure, along with a variation on the second. These are the sample interquartile range, the sample standard deviation and its close relation, the sample variance.

Measuring spread?

The sample interquartile range

The sample interquartile range is defined to be the difference between the two sample quartiles, that is, the sample upper quartile minus the sample lower quartile. As such, and as you have just seen in Activity 17, the sample interquartile range is the length of the box in a boxplot. This interpretation makes it clear that the interquartile range is measuring the amount of spread in the data, at least over its central portion, defined to be approximately its middle 50%.

The sample interquartile range

The **sample interquartile range** is defined as

$$q_U - q_L,$$

where q_U is the sample upper quartile and q_L is the sample lower quartile.

Example 18 *Calculating a sample interquartile range*

In Example 17, the quartiles for the percentages of adults who were members of sports clubs in six areas in England were calculated to be $q_L = 17.225$ and $q_U = 22.875$. So for these data,

$$q_U - q_L = 22.875 - 17.225 = 5.65.$$

That is, the sample interquartile range is approximately 5.7%.

Activity 18 *Calculating another sample interquartile range*

For the data on β endorphin levels of collapsed runners, calculate the sample interquartile range. Include the potential outlier in your calculations. (Remember that you obtained the sample quartiles for these data in Activity 16.)

The sample standard deviation

The sample standard deviation, like the sample mean, is a summary measure whose calculation involves all of the data values. At the heart of the calculation are 'deviations', the differences between the data points and their mean; these are $(x_i - \overline{x})$ for $x = 1, 2, \ldots, n$. It is these deviations that give the standard deviation its name. As the aim is to measure how spread out the data values are, regardless of whether they are above or below their mean, the deviations are first squared. These squared deviations are then averaged (although not quite in the usual way, as you will soon see). The final step, taking the square root of the averaged squared deviations, ensures that the standard deviation has the same units as the data values (for example, if the data are given in kilograms, then the sample mean and the sample standard deviation are given in kilograms also). Mathematically, this process is expressed as follows.

The same could be achieved by just ignoring the signs of the deviations. However, squaring deviations leads to a measure of spread that is easier to deal with mathematically.

The sample standard deviation

If the n values in a dataset are denoted $x_1, x_2, \ldots, x_n$ and their sample mean is

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

then the **sample standard deviation**, s, is defined by

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

In M248, the square root sign is always taken to mean the positive square root. For example, $\sqrt{9} = 3$ even though -3 is also a square root of 9.

As you will see later in M248, the division by $(n-1)$ rather than by n gives the sample standard deviation better statistical properties.

Note that the square of the standard deviation, s^2, is useful in its own right. So this quantity has its own name, the **sample variance**. Explicitly, the sample variance is given by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Example 19 *Calculating a sample standard deviation*

For the sample of areas in England introduced in Example 15, $n = 6$ and the mean percentage of adults who were members of sports clubs is $\overline{x} = 20.3\%$. So the sample variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{5}\sum_{i=1}^{6}(x_i - 20.3)^2$$
$$= \frac{1}{5}\left((19.1 - 20.3)^2 + (17.4 - 20.3)^2 + \cdots + (22.6 - 20.3)^2\right)$$
$$= \frac{1}{5}\left(1.44 + 8.41 + \cdots + 5.29\right) = \frac{43.66}{5} = 8.732.$$

This means that the sample standard deviation is $s = \sqrt{8.732} \simeq 2.95\%$.

Activity 19 *Calculating another sample standard deviation*

For the data on β endorphin levels of collapsed runners most recently given in Activity 16, calculate the sample standard deviation. Include the potential outlier in your calculations. To make the calculation a little less tedious, it is sufficient here to use the value of the sample mean that was calculated correct to one decimal place in Activity 12, namely, 138.6 pmol/l.

Interquartile range or standard deviation?

As with the measures of location, two (main) measures of spread have been described in this section: the sample interquartile range and the sample standard deviation. Before focusing on a comparison of the two, we'll include the sample variance in the considerations of the next activity.

Activity 20 *A negative spread?*

(a) Is it ever possible for the sample interquartile range to be negative?

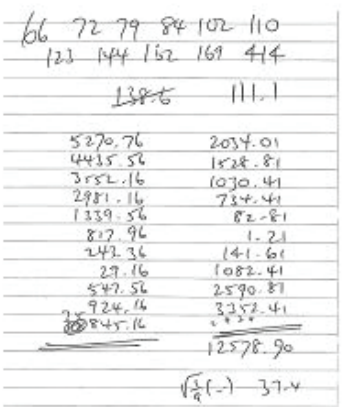
(b) Is it ever possible for the sample variance to be negative?

(c) Is it ever possible for the sample standard deviation to be negative?

So, as you argued in Activity 20, neither the sample interquartile range nor the sample standard deviation can ever be negative. In particular, this means that any claim that a sample interquartile range or a sample standard deviation (or a sample variance) is negative indicates that there has been an error in the calculation somewhere!

Both measures can take the value zero but these instances correspond to situations with little or no variability in the data, which are uninteresting statistically.

Furthermore, the more spread out the data are, the larger both quantities tend to be. In the case of the sample interquartile range, this is because spreading out the data results in the sample lower quartile and the sample upper quartile becoming further apart. In the case of the sample standard deviation, this is because spreading out the data results in the squared deviations from the mean becoming larger.



You may or may not have written so much out!

Although the sample interquartile range and the sample standard deviation share the above similarities, the values of the measures themselves are not directly comparable. That is, it is not possible to say that one dataset (say dataset 'A') is more spread out than another dataset (dataset 'B') just because the interquartile range for 'A' is more than the standard deviation for 'B'.

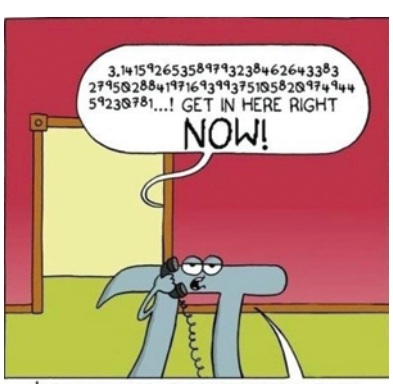
So, which one to choose?

The interquartile range is a resistant measure. For example, had we asked you to recalculate the sample interquartile range for the β endorphin levels of collapsed runners without the outlying value of 414, you would have found (using quartile values from the solution to Activity 16(b)), that $q_U - q_L = 148.5 - 77.25 = 71.25$ pmol/l. This is not *very* different from the value of 83 pmol/l that you found for the case with the outlier included in Activity 18. Further, the sample quartiles on which it is based share similarities with the sample median: the sample quartiles and the sample median are each based on picking out values from the dataset that are certain amounts of the way along the list of ordered data values. So the sample interquartile range is usually quoted when the sample median is a good choice of measure of location.

The standard deviation is based on deviations about the mean. It is not a resistant measure. For example, had we made you do a parallel tedious calculation to that of Activity 19 but without the outlying value 414 pmol/l, you would have found that the sample standard deviation reduces drastically from the 98.0 pmol/l you found there to 37.4 pmol/l. Again, however, as with the mean, other considerations can favour the standard deviation. So the sample standard deviation is usually quoted when the sample mean is a good choice of measure of location.

4.4 A note on accuracy

To what accuracy should you give the results of calculations? If you look through the examples in this section, you will find that, in general, results have been given either exactly or, when approximated, to the same accuracy as the data or rounded to one decimal place or one significant figure more than is given in the data. There is no hard-and-fast rule about what you should do: appropriate accuracy depends on a number of factors, including the reliability of the data and the size of the dataset. However, you should avoid rounding the data either too much, so that valuable information is lost, or too little, thus suggesting that your results are more accurate than can be justified from the available data.



I'VE GOT TO GO. MY MOM ONLY USES MY FULL NAME WHEN I'M IN BIG TOUBLE.

As a rough guide, it is usually satisfactory to round a result to one significant figure more than is given in the data. But note that this rough guide applies only to results quoted at the ends of calculations: intermediate results should not be rounded. If you round a result and then use the rounded value in subsequent calculations – for instance, if you use a rounded value for the mean when calculating the standard deviation of a dataset – then this sometimes leads to quite serious inaccuracies (known as rounding errors).

Statisticians are not obsessed with degrees of rounding. This is because any rounding errors are usually insignificant relative to other inaccuracies due to assumptions and approximations involved in the statistical modelling process.

4.5 Numerical summaries using Minitab

The work in this subsection consists of a chapter in Computer Book A, in which you will see how to use Minitab to calculate the numerical summaries that have been described in this section.

Refer to Chapter 3 of Computer Book A for the rest of the work in this section.

Exercises on Section 4

Exercise 7 *Quartiles of the chondrite data*

In Activity 14(b), you calculated the sample median for the dataset on the percentages of silica found in each of $n = 22$ chondrites given (in ordered form) in Table 10.

(a) Calculate the sample lower quartile and sample upper quartile of the percentages of silica in these chondrites.

(b) Hence calculate the sample interquartile range of the percentages of silica in these chondrites.

Exercise 8 *Location and spread of a sample of numbers of nuclear power stations*

In Table 1, the numbers of operational power stations in 2014 were given for all countries which had at least one such power station. From this population, a random sample of seven countries was then selected. The numbers of operational power stations for the selected areas are as follows.

99 23 1 3 15 4 9

(a) Calculate the sample median, quartiles and interquartile range of these data.

(b) Calculate the sample mean and standard deviation of these data.

(c) Give a brief summary of what parts (a) and (b) tell you about the location and spread of the numbers of operational power stations in the countries selected in this sample.

5 Comparing variables graphically

So far in this unit, you have been exploring techniques for getting a feel for data one variable at a time. In this section, the focus is on looking at two or more variables at the same time, with a view to comparing their distributions (in Subsections 5.1–5.3) or understanding the relationship between them (Subsection 5.4).

The graphical methods considered will be: side-by-side bar charts in Subsection 5.1; unit-area histograms in Subsection 5.2; comparative boxplots in Subsection 5.3; and scatterplots in Subsection 5.4. (Unit-area histograms can also be used for single datasets in place of frequency histograms.) Finally, in Subsection 5.5, you will use Minitab to obtain and investigate the graphs described in this section.

5.1 Side-by-side bar charts

When comparing different categorical or discrete variables, it is always possible to place bar charts next to one another; this way, by glancing between them, differences and similarities can be picked up. However, when the variables are actually measurements using the same categories but on a different group of objects, we can do better than that: the two bar charts can be merged so that the bars relating to the same categories are adjacent. Such a combined bar chart is known as a **side-by-side bar chart**.

Example 20 *Participants in a weight loss trial*

In Example 5, data from a weight loss clinical trial were introduced. A side-by-side bar chart giving the number of male and female participants in the trial is shown in Figure 19. (Here, the data from all 83 participants are included, not just the 10 listed in Table 5.)

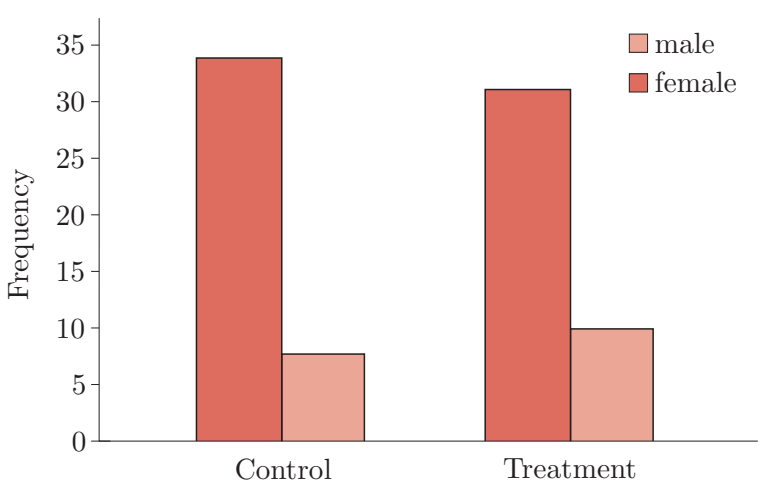


Figure 19 Numbers of females and males in each of the two treatment groups of a weight loss trial

In this side-by-side bar chart, there are two pairs of bars displayed: the left-hand pair represents the data for the control group, and the right-hand pair the data for the treatment group. Within each pair, there is a bar for the number of females and a bar for the number of males.

The heights of the bars representing the numbers of females are similar in height, indicating that the numbers of females in the control and treatment groups were similar. The same can be said for the numbers of males in the two groups. However, in both pairs of bars, it is clear that the bars representing females are much taller than the bars representing males. So in this weight loss clinical trial, there were many more female participants than male participants. The ratio of female to male participants in the two groups was broadly similar (as reflected in the heights of the bars within treatment groups), with a slightly lower ratio of females to males being apparent in the treatment group compared with the control group.

Activity 21 *Gender distribution of the UK workforce*

In Activity 7, you considered a bar chart of the distribution of the total UK workforce at the end of 2015. The data in Table 2, given in Example 2, break down the workforce numbers for males and females separately. A side-by-side bar chart of these data is given in Figure 20.

On the right, a male in a 'Caring & leisure' occupation

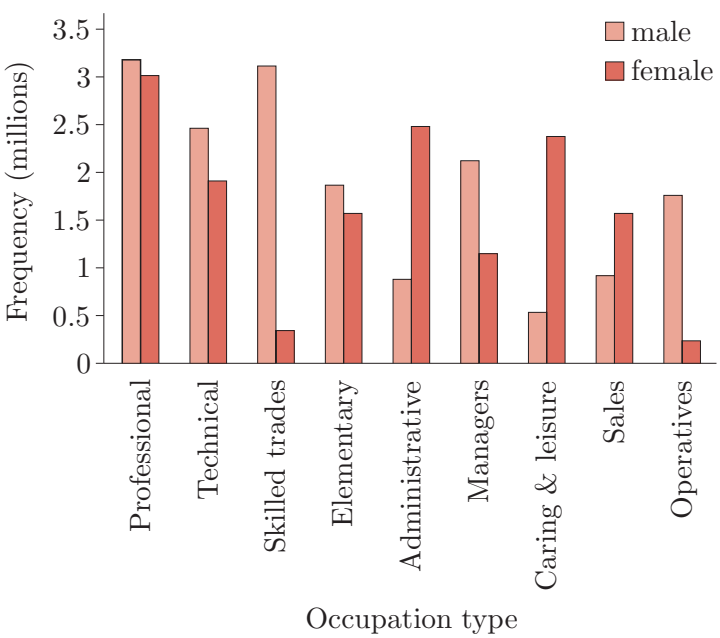


Figure 20 UK workforce, October to December 2015, by gender

Using this side-by-side bar chart, comment on how the numbers employed in each occupation type compare for males and females.

5.2 Unit-area histograms

This applies to bar charts too, as you saw in Example 20.

As with bar charts, the most straightforward way of comparing histograms is to place histograms for two different variables next to one another. However, one consequence of doing this – if the histograms are plotted on the same vertical scale – is that the most obvious feature that might show up is simply that one sample size is larger than the other. When the sample sizes are very different, this could obscure comparisons of other aspects of the shapes of the distributions of the variables.

Example 21 *Comparing weight changes with frequency histograms*

In Figure 21, two frequency histograms are given. Both depict data from the weight loss clinical trial described in Example 5 relating to the change in weight of participants after two weeks. The histogram in Figure 21(a) displays the data for females in the control group, while the histogram in Figure 21(b) displays the data for males in the control group.

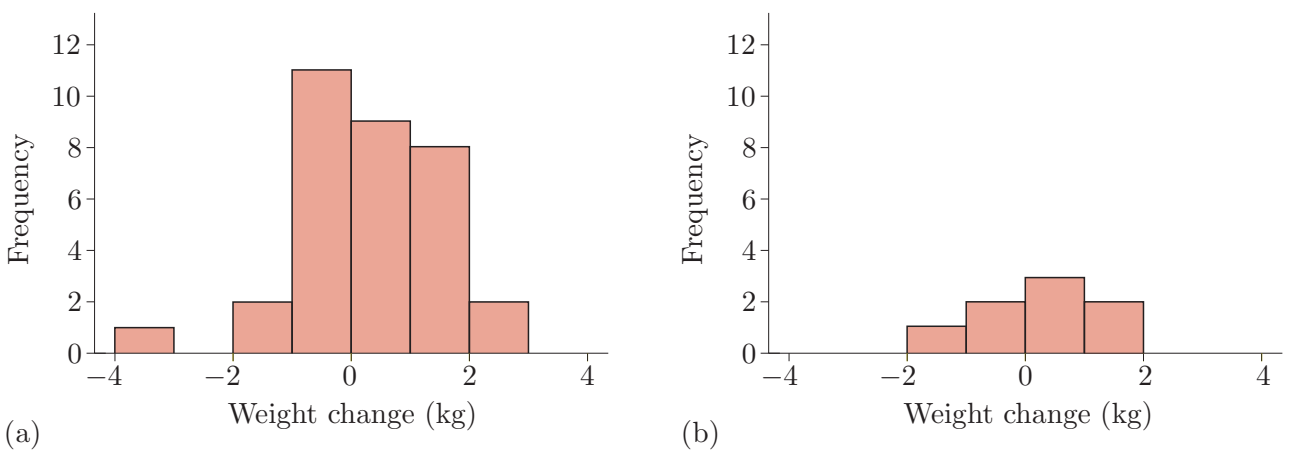


Figure 21 Frequency histograms of weight change for (a) females and (b) males in the control group

The bars on the histogram for females are generally much longer than the bars on the histogram for males, reflecting the fact that the control group consisted of many more females than males.

This feature can be removed by plotting *unit-area histograms*. Instead of plotting frequencies for each group, *scaled* frequencies are plotted.

Unit-area histograms

A **unit-area histogram** is a frequency histogram in which the frequencies are scaled so that the total area of the bars in the histogram is 1.

Before applying this notion to Figure 21, let us look at the effect of making a single frequency histogram into a unit-area histogram.

Example 22 *A unit-area histogram of sports club membership*

In Figure 4 of Example 9, a frequency histogram of the sports club membership data was given. It is repeated here for convenience.

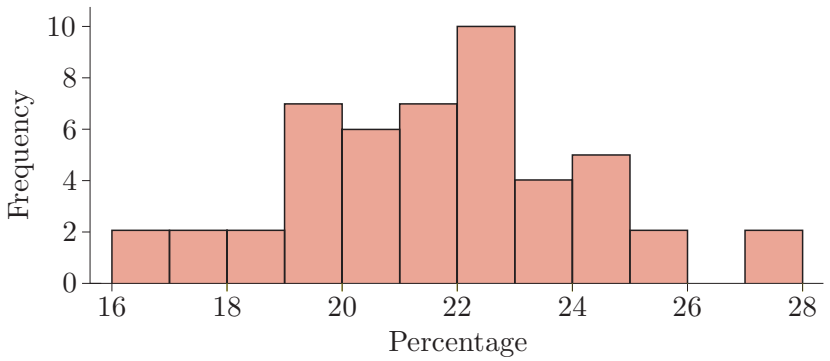


Figure 22 A frequency histogram of the sports club membership data

A unit-area histogram of the same data, using the same cutpoints, is shown in Figure 23.

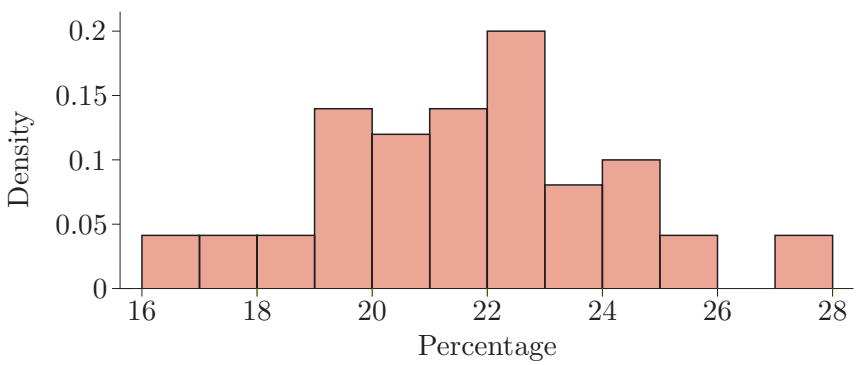


Figure 23 A unit-area histogram of the sports club membership data

If you are thinking that the only difference between Figures 22 and 23 is that the vertical scale has changed, you'd be right. The *shape* of the histogram has remained unaltered. This small change is all that is needed to ensure that the total area of the bars takes the value 1. You will see how this works in the following activity.

Note that the label 'Density', replacing 'Frequency', on the vertical axis indicates that a unit-area histogram is being plotted.

Activity 22 *Total area of bars*

Consider the *frequency* version of the histogram of sports club membership given in Figure 22. Numbering the bars from left to right, the frequencies represented by each bar are given in the following table. (You will fill in the blank row in the table shortly.) There's a zero frequency between what we've called bars 10 and 11.

Table 11 Frequencies in Figure 22

Bar number	1	2	3	4	5	6	7	8	9	10	11
Frequency	2	2	2	7	6	7	10	4	5	2	2
Scaled area											

(a) The area of a single bar is its height times its width, which in Figure 22 is the frequency associated with the bar times 1 (since the width of each bar in this case happens to be 1; usually, it isn't). Calculate the total area of the bars in Figure 22.

(b) In order to scale the frequencies so that the total area of the bars is 1, divide the frequency of each bar by the total area of the bars to obtain the scaled frequencies. Again, these scaled frequencies are scaled areas since the width of each bar is 1. Round these values correct to two decimal places and put them in the blank row in Table 11. Check that their sum is 1.

(c) Check visually that the areas that you calculated in part (b) approximately match the heights of the unit-area histogram in Figure 23. (Again, in this case, areas equal heights because bar widths are 1.)

The approximation here derives from the rounding.

It might seem a trivial change to change the vertical scale so that the total area is equal to 1. However, by doing so, it allows us to focus on comparing the shapes of the histograms.

Example 23 *Comparing weight changes with unit-area histograms*

In Example 21, frequency histograms of weight changes over two weeks for participants in the control group of a weight loss clinical trial were given; Figure 21(a) is a frequency histogram for females, Figure 21(b) for males. Unit-area histograms for these data are given in Figure 24.

The unit-area histograms in Figure 24 are fairly similar. The bars are a little higher and more closely packed for males compared with females. This indicates that the weight changes were less spread out for males compared with females. Otherwise, both histograms indicate just one mode and look approximately symmetric.

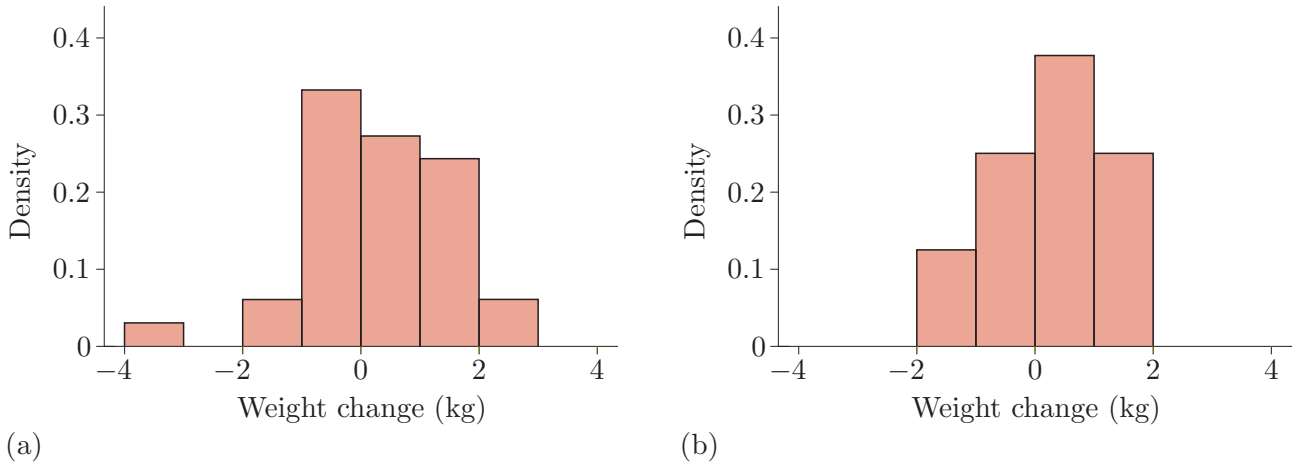


Figure 24 Unit-area histograms of weight change for (a) females and (b) males in the control group

Activity 23 *Recall of memories*

In a study of memory recall times, a series of stimulus words was shown to a subject on a computer screen. For each word, the subject was instructed to recall either a pleasant or an unpleasant memory associated with that word. Successful recall of a memory was indicated by the subject pressing a bar on a computer keyboard. Of key interest in this study was whether pleasant memories could be recalled more easily and quickly than unpleasant ones. Unit-area histograms for the two samples are shown in Figure 25.

Use the unit-area histograms to compare the distributions of recall times for the two types of memory.

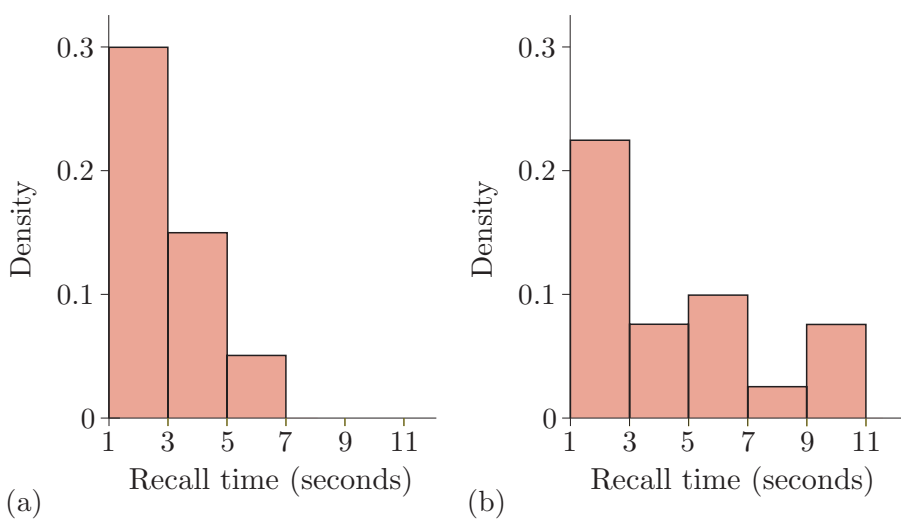


Figure 25 Recall times for (a) pleasant and (b) unpleasant memories

(Source: Dunn, G. and Master, D. (1982) 'Latency models: the statistical analysis of response times', *Psychological Medicine*, vol. 12, pp. 659–65)

You have seen how unit-area histograms make it easier to compare data from two groups, particularly when the sample sizes are very different. Another advantage of unit-area histograms, which will not be investigated here, is that they are appropriate when the widths of bins are not the same. They also have an important role to play in linking the distribution of the data with models for the distribution of the data, as you will see in Unit 2. It is there that the reason for the mysterious label 'density' will become clear.

5.3 Comparative boxplots

It is not really feasible to draw two histograms on the same plot. An advantage of using boxplots to display the distribution of a continuous variable is that plotting two, or even more, boxplots on the same diagram is straightforward. The boxplots are just stacked one above the other. Such plots involving multiple boxplots are known as **comparative boxplots**.

Example 24 *A comparative boxplot of weight change*

In Example 23, you looked at unit-area histograms of weight change two weeks into a weight loss clinical trial. The unit-area histograms displayed the data for the females in the control group and the males in the control group separately. A comparative boxplot of the same data is given in Figure 26.

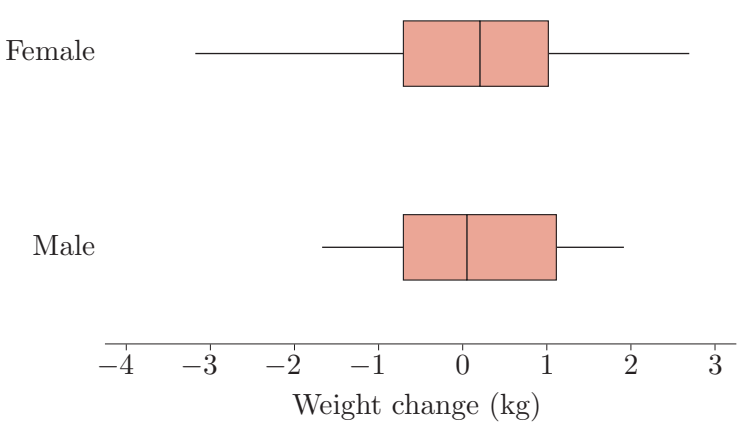


Figure 26 Comparative boxplot of weight change in the control group

In Figure 26, the box parts of both boxplots are very similar. This indicates that the central 50% of the data for females and for males are very similar. However, away from the central 50%, the data for the females are more spread than for the males, the whiskers being longer.

Activity 24 *A comparative boxplot of memory recall data*

In Activity 23, you used unit-area histograms to compare the distributions of recall times for pleasant and unpleasant memories. The comparative boxplot shown in Figure 27 represents the same data.

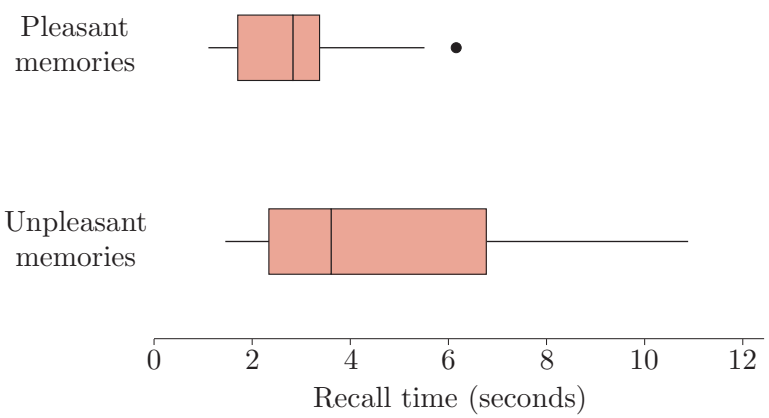


Figure 27 Comparative boxplot of memory recall times

(a) In the solution to Activity 23, it was noted that the distribution of recall times of unpleasant memories was more spread out than the distribution of pleasant memories. Is this backed up by Figure 27? Why or why not?

(b) Also in the solution to Activity 23, it was noted that both distributions appear to be right-skew. Is this also backed up by Figure 27? Why or why not?

(c) Identify one other difference between the two distributions that is apparent in Figure 27.

Comparative boxplots have an arguably even more important role in comparing the distributions of more than two variables. You will explore such a situation in your computer work in Subsection 5.5.

5.4 Scatterplots

So far, all the comparative diagrams have assumed that the variables are not linked, that is, that the measurements were made on different cases. When two variables are linked, a **scatterplot** is used instead.

The variables are often continuous but may be discrete.

On a scatterplot, each case is represented by a point. The position of this point along the horizontal (x) axis indicates the value of one variable for this case, while the position of the point along the vertical (y) axis indicates the value of the other variable for the same case.

Example 25 *Defects in the Trans-Alaska oil pipeline*

This example concerns defects in the Trans-Alaska oil pipeline. Depths of defects were measured in the field using ultrasonic measuring equipment. The depths of those same defects (the cases) were later remeasured in a laboratory. There were $n = 107$ such linked pairs of measurements obtained in this way in an attempt to understand how well calibrated in-field measurements of pipeline defects were. The data are displayed on a scatterplot (Figure 28). (Unfortunately, units of measurement are not available.)

There are 107 points on this scatterplot, each point representing a defect in the pipeline. For example, the topmost point, well to the right-hand side of the plot, corresponds to a pipeline defect that was measured in the field to be at a depth of 85 but in the laboratory to be at a depth of 80.

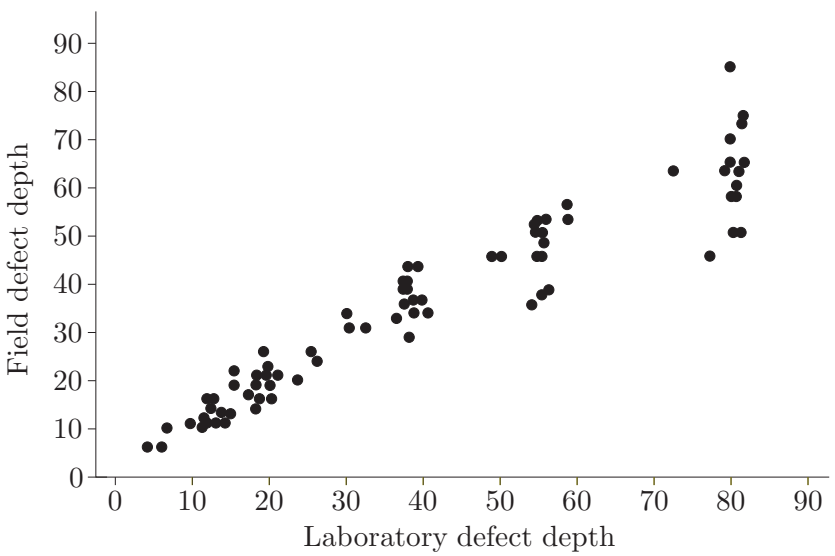


Figure 28 Field and laboratory defect depths in the Trans-Alaska oil pipeline

(Source: data collected by Harry Berger, US National Institute of Standards and Technology)

When interpreting scatterplots, the focus is on what the pattern of points suggests about the relationship (or association) between the two variables. Things to consider when interpreting a scatterplot include those on the following checklist.

Scatterplot interpretation checklist

- Is the relationship positive, negative or neither?
- Is the relationship linear or non-linear?
- Is the relationship strong or weak?
- Are there any outliers?

Example 26 *Interpreting a scatterplot*

So what does Figure 28 tell us about the relationship between pipeline defect depth measurements as made in the field and those made in the laboratory?

- First, it suggests that there is a positive relationship between the measurements, because as one measurement takes larger values, so does the other.
- Second, the relationship appears to be roughly linear because the main upward trend in the data, broadly speaking, follows a straight line. (Alternatively, you might perceive a hint of non-linearity, perhaps arguing that the trend 'curves over' for larger values of the measurements.)
- Third, the relationship appears to be at least moderately strong: points generally lie fairly close to any 'central' trend line, although the spread of points about the trend is not *very* narrow.
- And fourth, there are no clear outliers, that is, no points very different from the rest, although one might be concerned about points like that singled out in Example 25 because its field measurement seems further out of line with its laboratory measurement than is the case for most other points.

In this example, an extra feature that you might perceive in Figure 28 is that the amount of spread of the points about any central line appears to increase as the values of the measurements increase.

As with the interpretation of other graphical output, scatterplot interpretations are somewhat subjective, with different people sometimes arriving at different – tentative – conclusions based on the same plot.

Activity 25 *Membership of, and coaching in, sports clubs*

In Example 4, data giving the percentages of adults who were members of sports clubs in 49 areas in England in 2014–15 were introduced. In fact, also available for these areas in 2014–15 are the percentages of adults who received coaching in a sport in the previous year. These data can be displayed on a scatterplot such as that in Figure 29 (overleaf).

Using the interpretation checklist, describe the relationship between the percentage of adults who were members of sports clubs and the percentage who received sports coaching in the previous year.

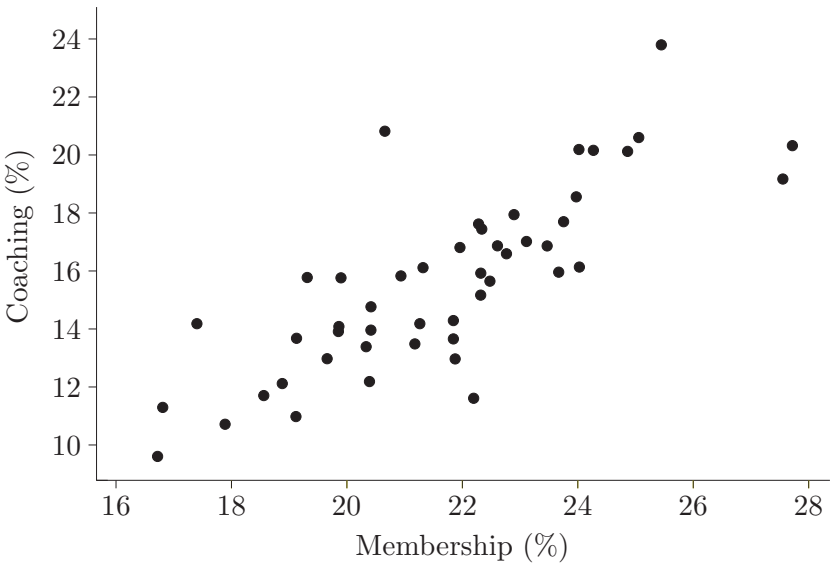


Figure 29 Sports clubs membership and sports coaching in sports partnership areas

5.5 Comparative plots using Minitab

The work in this subsection consists of another chapter from Computer Book A. You will see how to produce the plots introduced in this section: side-by-side bar charts, unit-area histograms, comparative boxplots and scatterplots.

Refer to Chapter 4 of Computer Book A for the rest of the work in this section.

Exercises on Section 5

Exercise 9 *Deaths from high blood glucose*

The percentages of deaths between ages 20 and 69 attributable to high blood glucose across the world in 2012 are plotted in Figure 30, categorised by the income level of the country and by males and females. (In this context, percentages are just another rescaling of frequencies, removing the effects of total numbers of people in each category.)

Use the side-by-side bar chart in Figure 30 to describe briefly the differences between countries with different income levels and between males and females.

To quote from the source of the data: 'the number of adults living with diabetes has almost quadrupled since 1980 to 422 million'.

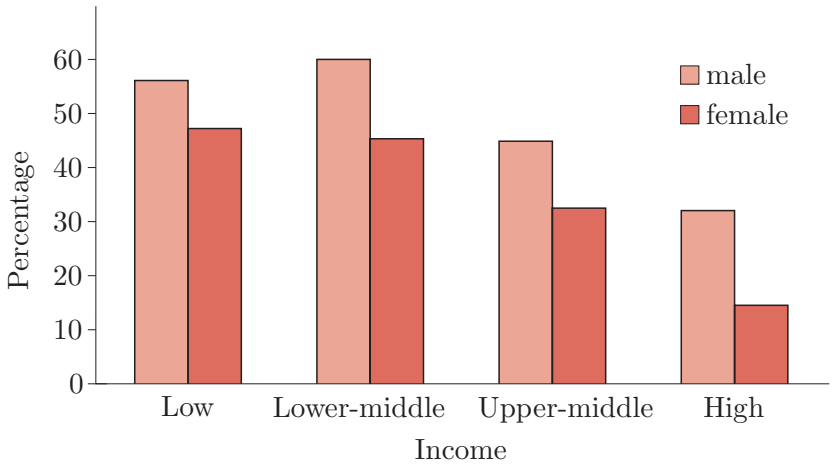


Figure 30 Percentages of deaths attributable to high glucose levels

(Source: World Health Organization (2016) *Global Report on Diabetes*)

Exercise 10 *Family sizes*

Discrete data are often treated as if they are continuous data, especially when the data can take on a high number of different values.

In Example 3, data from the 1941 Canadian census were introduced on the numbers of children born to mothers satisfying particular religious, age and location specifications, and who had been educated for seven years or more. In fact, a similar dataset was also produced on the numbers of children born to mothers satisfying the same specifications, except that they had been educated for less than seven years.

Figure 31 displays a comparative boxplot of both datasets, while in Figure 32 (overleaf), unit-area histograms of the same data are shown.

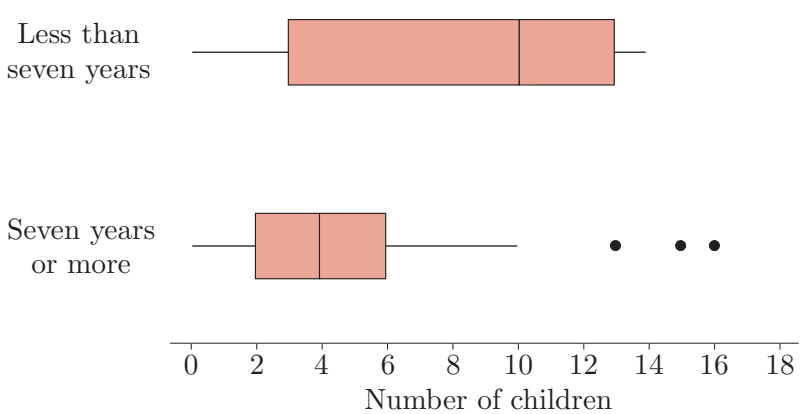


Figure 31 Boxplots of numbers of children born

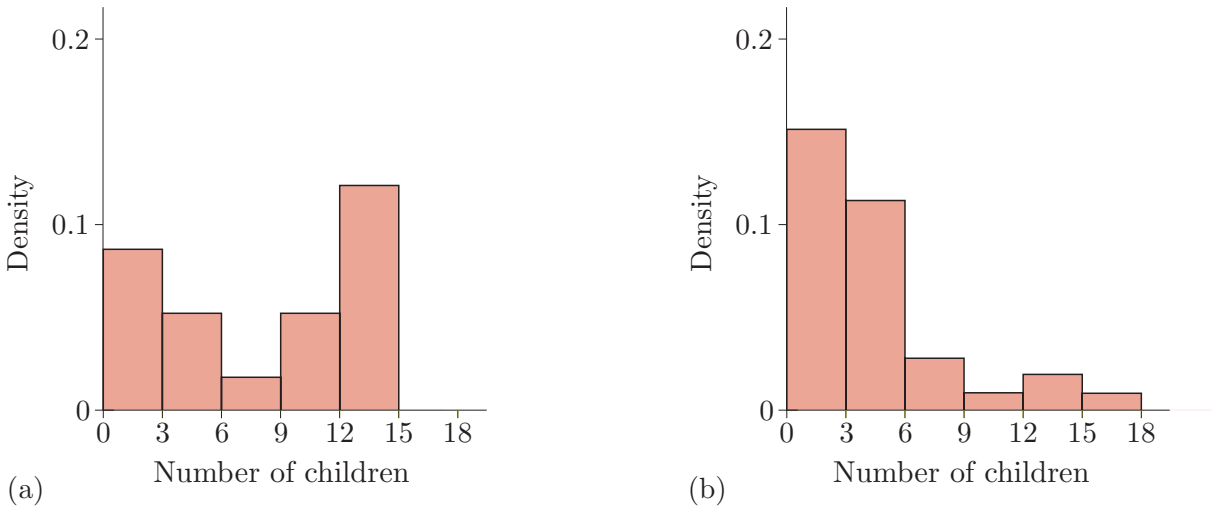


Figure 32 Unit-area histograms of numbers of children born to mothers educated for (a) less than seven years, (b) seven years or more

(a) Describe one feature of the data that is apparent in both the comparative boxplot and the unit-area histograms.

(b) Describe one feature of the data that is apparent from the unit-area histograms that is not apparent in the comparative boxplot.

(c) Describe one feature of the data that is apparent from the comparative boxplot that is less obvious in the unit-area histograms.

Exercise 11 *Magnitude and depth of earthquakes*

In Exercise 6, data on earthquakes occurring in one 24-hour period were introduced. In that exercise, only the magnitudes of the earthquakes were considered. However, linked data on the depths of the earthquakes (in km) were also recorded. Figure 33 is a scatterplot of the magnitudes of the earthquakes against their depths.

Describe the relationship, if any, between the magnitudes and depths of earthquakes that this scatterplot suggests.

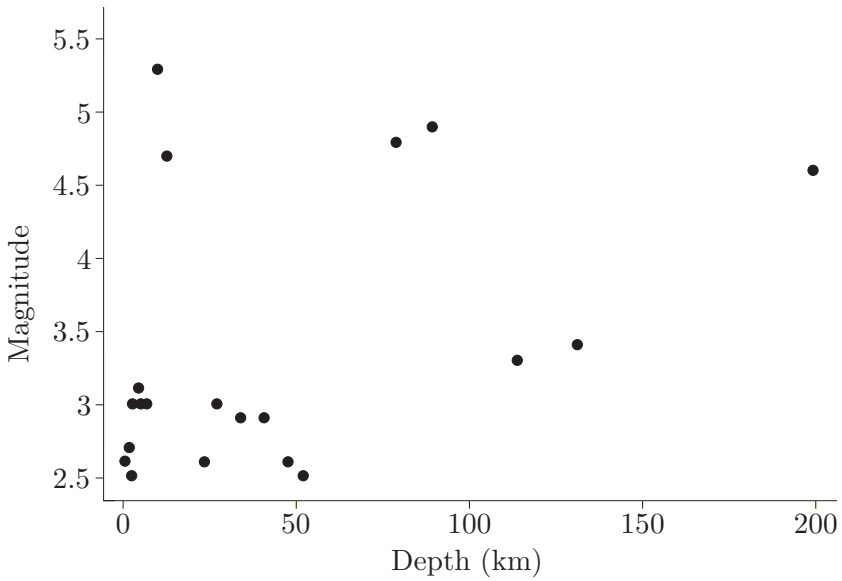


Figure 33 Magnitudes and depths of earthquakes

6 Unit 1 and the rest of the module

Unit 1 has a different 'feel' to what much of the rest of this module will have, for two main reasons.

The first of these is that its purpose has largely been revision of ideas and techniques, most of which you should have come across before embarking on M248. At other points in the module, other notions and methods will arise that you might also have met before, and which will be revised as and when necessary. However, most of the rest of the module will, of course, be expected to be new to you (including deeper consideration of some ideas whose basics you might have met before).

The second reason is that Unit 1 has concentrated on looking at data in a descriptive fashion, in order to get some idea of its features and foibles through graphical exploration and some simple numerical summaries. The remainder of M248, and indeed of much of statistics, is concerned with proceeding on to developing appropriate *models* for data, and with using those models to help us answer real, practical, questions to which the data have something relevant to contribute. The latter process is known as *statistical inference*. Graphs, and especially numerical summaries, but of models as well as of data, will continue to play an enormous role in the module.

Summary

In this unit, you have been (re)introduced to a number of ways of representing data graphically and of summarising data numerically. We began by looking at some datasets and considering informally the kinds of questions they might be used to answer. The distinction between populations and samples has also been stressed.

An important first stage in any assessment of a collection of data is to represent the data, if possible, in some informative diagrammatic way. Useful graphical representations that you have met in this unit include bar charts, histograms, boxplots and scatterplots. Bar charts are generally used with categorical data, or with numerical data that are discrete (counted rather than measured); side-by-side bar charts can be used to display more than one such variable. Histograms and boxplots are generally used with continuous (measured) data. Histograms come in frequency and unit-area versions which differ only in their 'vertical' scaling; a comparative boxplot allows more than one continuous variable to be displayed at the same time. Histograms need a reasonably large dataset and are sensitive to the choice of cutpoints; boxplots cannot show how many modes a distribution has. Scatterplots are used to investigate the relationship between two numerical variables (which are often continuous but may be discrete).

Features of the distribution of a dataset have been described. Modes and modality (unimodal, bimodal or multimodal?) are one important aspect; skew, which might be right-skew or left-skew, is another.

Numerical summaries of data are also very important. Two main pairs of summaries for assessing location and spread have been described. The measures of location that have been discussed are the sample mean and the sample median, and the measures of spread are the sample standard deviation (together with a closely related measure, the sample variance) and the sample interquartile range. Sample lower and upper quartiles are required for the interquartile range and for boxplots. The median and interquartile range are more resistant to unusual values in the data than are the mean and standard deviation, although the mean and standard deviation also have properties which make them good to work with.

The data analysis software Minitab has been introduced. As well as learning the basics of using the software, you have seen how to use it to produce all of the graphical displays mentioned above and to calculate the numerical summaries.

Learning outcomes

After you have worked through this unit, you should be able to:

- appreciate that many datasets can be thought of as being a representative sample from an underlying population
- gain an overall 'feel' for data and the way they are distributed by constructing appropriate graphical displays and calculating numerical summaries
- recognise that producing an appropriate graphical representation or appropriate numerical summaries can help greatly in comparing sets of data
- understand that categorical and discrete data can be represented in bar charts, and that when interest lies in comparing two or more such datasets, a side-by-side bar chart can be useful
- appreciate that in order to examine a continuous variable, both histograms and boxplots are appropriate graphics
- understand the difference between a frequency histogram and a unit-area histogram
- recognise that a boxplot shows less detail about a distribution than a histogram or bar chart, but that boxplots are particularly useful for comparing two or more datasets
- use scatterplots when the values of two linked numerical variables are available and the aim is to investigate the relationship between the two variables
- calculate the sample mean and sample median as measures of location of a dataset
- calculate the sample standard deviation, sample variance and sample interquartile range as measures of spread of a dataset, the last named requiring the calculation of sample lower and upper quartiles
- appreciate that the sample median and interquartile range are more resistant measures than are the sample mean and standard deviation
- use some of the basic facilities of Minitab, especially to obtain the graphical representations and numerical summaries listed above.

Solutions to activities

Solution to Activity 1

The structures of the datasets are similar in that they each contain information about a set of entities, objects or individuals: countries in the case of Table 1, occupation types in the case of Table 2, mothers satisfying particular criteria in Table 3, areas in England in the case of Table 4, participants in a clinical trial in the case of Table 5, and patients in the case of Table 6. Furthermore, for each dataset, one or more characteristics are known about each object. For example, in Table 5, there are four characteristics associated with each person: the treatment group they're in, the number of training sessions attended, their gender and their weight change over two weeks.

In Table 2, one characteristic happens to be the sum of the other two.

However, the numbers of objects and the numbers of characteristics given in each dataset are not the same. Table 1 provides information about 30 objects, and information is provided on just one characteristic for each. Table 2 provides information about 9 objects and 3 characteristics. Table 3 provides information about 35 objects and 1 characteristic. Table 4 provides information about 49 objects and again just 1 characteristic. Table 5 provides information about 10 objects and 4 characteristics, the complete dataset having included information about 83 objects altogether. (And in the complete clinical trial, very many more characteristics were recorded about the participants, too.) Table 6 provides information about 55 objects and 5 characteristics.

Also, for many of the characteristics the entries are numerical, but for other characteristics the entries are not numerical. For example, for the data given in Table 5, the number of training sessions attended and weight change are numerical, but the groups into which patients are put and their gender are not numerical.

Solution to Activity 2

(a) Discrete. The number of operational power stations is a count, so only particular (integer) numbers make sense. For example, while a country might have, say, 3 operational power stations, it does not make sense to say that a country has 3.1755 operational power stations.

(b) Continuous. The percentage of adults can be thought of as a measurement which can take any value (to any number of decimal places) between 0 and 100.

(c) Nominal. The values 'A' and 'B' are just labels associated with the two methods under consideration.

(d) Ordinal. The possible values 'small', 'medium' and 'large' are labels, but with the categories having a natural ordering with respect to how big the tattoo was.

(e) Ordinal. Although numbers have been used to indicate the degree of success of the tattoo removal, the values 1, 2, 3 and 4 are really just labels. There is, however, a natural ordering to the categories with respect to how successful the removal was.

(f) Strictly speaking, as a count, this variable is a discrete variable. However, in practice, this variable would usually be treated as a continuous variable. This is because the numbers concerned are so large that if we were to plot out all the possible numbers we could get, any gaps between numbers would be so small as to be unnoticeable.

Solution to Activity 3

(a) It is reasonable to treat all three variables as continuous; β endorphin concentrations appear to be capable of taking any of a wide range of positive values.

(b) The variables in the first two columns of Table 7, the blood plasma β endorphin concentrations in each 'normal runner before race' and 'same runner after race', are linked as they relate to measurements on the same individuals in a group of 11 runners. However, the variable representing the blood plasma β endorphin concentrations in each 'collapsed runner' is not linked to the other two variables as the collapsed runners were different individuals to the 'normal' runners.

Solution to Activity 4

(a) Population. All people in employment in the UK in the last quarter of 2015 are included in the dataset. (At least, that is the aim of the Office for National Statistics, who compiled the data.)

(b) Sample. It might be hoped that the results for participants in the clinical trial would generalise to perhaps the entire population of people interested in achieving loss of weight.

(c) Population. The 49 areas cover the whole of England. So there are no areas left out of the dataset.

(d) Sample. Reasonable suggestions for the underlying population include: all runners in that particular Great North Run; all runners in the Great North Run, regardless of year; all runners competing in races of similar distance to the Great North Run.

(e) Sample. It is reasonable to regard all the adults in England and Wales in 2015–16 as the underlying population. The Crime Survey for England and Wales involves tens of thousands of participants, yet its output remains a very large sample rather than a population (which in this case is of the order of a thousand times bigger).

Solution to Activity 5

This dataset could represent either a population or a sample, depending on how the underlying population is defined. If the underlying population is taken to be earthquakes of magnitude at least 2.5 that occurred during a 24-hour period that was, for some reason, of particular interest, then the data would constitute a population. However, there are many other 24-hour periods that could have been chosen. So, regarding the underlying population as being all earthquakes of magnitude at least 2.5, these data form a sample.

Solution to Activity 6

(a) No, this is not likely to be representative. The areas in South East England are likely to be more similar to each other than to areas in the rest of England. So by just concentrating on these areas, the full range of variation in sports club membership across England is likely to be missed.

(b) Yes, this is likely to be representative as it would be a simple random sample.

(c) Yes, this is likely to be representative. Even though the sample would be chosen in a systematic way, it is with respect to something that is not likely to be related to sports club membership.

Solution to Activity 7

The occupation types with the most workers are 'Professional' followed by 'Technical'. The occupation types with the fewest workers are 'Operatives' followed by 'Sales'. All the other categories have similar, intermediate numbers of workers.

Solution to Activity 8

(a) **Table 12**

Bin	Values	Frequency
1	16.0–18.0	4
2	18.0–20.0	9
3	20.0–22.0	13
4	22.0–24.0	14
5	24.0–26.0	7
6	26.0–28.0	2

(b) **Table 13**

Bin	Values	Frequency
1	16.0–20.0	13
2	20.0–24.0	27
3	24.0–28.0	9

(c) **Table 14**

Bin	Values	Frequency
1	15.0–17.0	2
2	17.0–19.0	4
3	19.0–21.0	13
4	21.0–23.0	17
5	23.0–25.0	9
6	25.0–27.0	2
7	27.0–29.0	2

Solution to Activity 9

(a) Figures 4, 5(a) and 5(b) are based on the same starting position as one another, namely, 16. (Figure 5(c) starts at 15, while Figure 6 starts at 16.5.)

(b) The bin width is smallest in Figure 4 (where it is 1), larger in Figure 5(a) (where it is 2) and larger again in Figure 5(b) (where it is 4). Figure 4 shows the most detail in the distribution of sports club membership, Figure 5(b) the least. The bin width controls the degree of detail shown in the histogram.

(c) Figures 4 and 6 are based on the same bin width as one another, namely, 1; Figures 5(a) and 5(c) are also based on the same bin width as one another, namely, 2. (Figure 5(b) uses bin width 4.)

(d) Comparing Figure 4 with Figure 6, we see a similar degree of 'bumpiness' but with bumps moved around in one figure relative to the other. Figure 5(a) gives the impression, perhaps, of a 'fatter' centre to the distribution and less in the way of 'tails' compared with Figure 5(c). Whatever the specific effect of starting position, it certainly seems to be quite considerable.

Solution to Activity 10

(a) Overall, the histogram appears to be unimodal with a peak around 77.5 years. That said, there is actually a second peak at about 67.5 years, but the corresponding bar is only very slightly higher than the one to its right, and most statisticians would discount it and claim this histogram to be unimodal.

The distribution is also left-skew. If a country had a life expectancy below 75 years, it could be far below. However, if a country had a life expectancy above 80 years, it could not be so far above 80 years.

(b) There are two clear peaks, one just above 50 minutes and another just below 80 minutes. This histogram is clearly bimodal: most often you are likely to have to wait well over an hour for the next eruption of Old Faithful, but there is also a good chance that your wait will be a little under an hour.

Solution to Activity 11

About 50% of the values lie between approximately 80 and 160 pmol/l, with most of the rest of the data between approximately 60 and 170 pmol/l. There is one outlier with a β endorphin concentration of over 400 pmol/l. In one way, this individual is of particular interest: what caused such a high β endorphin concentration? is it real or perhaps an error in measurement?

On the other hand, we might be more interested in the distribution of β endorphin concentrations for the more typical collapsed runners. Then, ignoring the outlier, the distribution of the β endorphin concentrations still does not seem to be symmetric. While the left whisker is longer than that on the right, the difference between the lengths of the whiskers is small. However, the right half of the box is longer than the left half, so there is some indication of right-skew in the data.

This is a small dataset, comprising just 11 measurements, so we should not over-interpret the data.

Solution to Activity 12

The sample mean of these concentrations is

$$\overline{x} = \frac{66 + 72 + 79 + 84 + 102 + 110 + 123 + 144 + 162 + 169 + 414}{11}$$
$$= \frac{1525}{11} \simeq 138.6.$$

The issue of how many decimal places to retain in calculations will be discussed later, in Subsection 4.4.

Thus the sample mean β endorphin concentration is approximately 138.6 pmol/l.

Solution to Activity 13

(a) In ascending order, the percentages are as follows:

$$x_{(1)} = x_5 = 16.7, \quad x_{(2)} = x_2 = 17.4, \quad x_{(3)} = x_1 = 19.1,$$
$$x_{(4)} = x_4 = 22.3, \quad x_{(5)} = x_6 = 22.6, \quad x_{(6)} = x_3 = 23.7.$$

(b) (i) $x_{(1)}$ is the smallest observation, that is, $x_{(1)} = x_5 = 16.7$.

(ii) As there are 6 observations, $x_{(6)}$ is the largest observation, that is, $x_{(6)} = x_3 = 23.7$.

(iii) $x_{(4)} = x_4 = 22.3$.

Solution to Activity 14

(a) There are $n = 11$ observations, so $\frac{1}{2}(n+1) = \frac{12}{2} = 6$ and hence $m = x_{\left(\frac{1}{2}(n+1)\right)} = x_{(6)}$. Notice that the concentrations given in Activity 12 are already in ascending order. So $x_{(6)} = x_6$ and $m = x_{(6)} = 110$. Thus the sample median β endorphin concentration is 110 pmol/l.

(b) For this sample, $n = 22$ so $\frac{1}{2}(n+1) = \frac{23}{2} = 11\frac{1}{2}$ and hence $m = x_{\left(\frac{1}{2}(n+1)\right)} = x_{\left(11\frac{1}{2}\right)}$. This means that m is the value halfway between $x_{(11)}$ and $x_{(12)}$. Again, the dataset is already in ordered form,

so

$$m = \frac{1}{2}\left(x_{(11)} + x_{(12)}\right) = \frac{1}{2}(28.69 + 29.36) = \frac{58.05}{2} = 29.025.$$

Thus the sample median percentage of silica is 29.025%.

Solution to Activity 15

(a) The mean for the 10 observations that are not outliers is

$$\begin{aligned}\overline{x} &= \frac{66 + 72 + 79 + 84 + 102 + 110 + 123 + 144 + 162 + 169}{10} \\ &= \frac{1111}{10} = 111.1.\end{aligned}$$

So the mean β endorphin concentration is now 111.1 pmol/l.

(b) Without the outlier there are 10 observations, so $n = 10$ and hence $m = x_{\left(\frac{1}{2}(n+1)\right)} = x_{\left(5\frac{1}{2}\right)}$. This means the median is halfway between $x_{(5)}$ and $x_{(6)}$. That is, $m = \frac{1}{2}(102 + 110) = 106$ pmol/l.

(c) Removing the outlier has changed the mean from 138.6 pmol/l to 111.1 pmol/l, whereas the median has only changed from 110 pmol/l to 106 pmol/l. This shows that the median is a resistant measure: its value does not depend very much on whether the outlier is included in the calculation or not. In contrast, whether or not the outlier is included makes a big difference to the value of the mean, showing that it is not a resistant measure.

Solution to Activity 16

(a) When all the observations are included, $n = 11$. So

$$q_L = x_{\left(\frac{1}{4}(n+1)\right)} = x_{\left(\frac{12}{4}\right)} = x_{(3)} = 79$$

and

$$q_U = x_{\left(\frac{3}{4}(n+1)\right)} = x_{\left(\frac{36}{4}\right)} = x_{(9)} = 162.$$

So when the outlier is included, the sample lower and upper quartiles are 79 pmol/l and 162 pmol/l, respectively.

(b) When the outlier is excluded, $n = 10$. So

$$\begin{aligned}q_L &= x_{\left(\frac{1}{4}(n+1)\right)} = x_{\left(\frac{11}{4}\right)} = x_{\left(2\frac{3}{4}\right)} \\ &= x_{(2)} + \frac{3}{4}\left(x_{(3)} - x_{(2)}\right) = 72 + \frac{3}{4}(79 - 72) = 77.25\end{aligned}$$

and

$$\begin{aligned}q_U &= x_{\left(\frac{3}{4}(n+1)\right)} = x_{\left(\frac{33}{4}\right)} = x_{\left(8\frac{1}{4}\right)} \\ &= x_{(8)} + \frac{1}{4}\left(x_{(9)} - x_{(8)}\right) = 144 + \frac{1}{4}(162 - 144) = 148.5.\end{aligned}$$

This means that when the outlier is excluded, the sample lower and upper quartiles are 77.25 pmol/l and 148.5 pmol/l, respectively.

(c) When the outlier is dropped from the calculation, the lower quartile does not change very much, while the upper quartile changes more than the lower quartile.

Solution to Activity 17

The lower limit of the box is the sample lower quartile. So approximately 25% of the data lie to the left of this point (either along the whisker or marked as potential outliers). Similarly, the upper limit of the box is the upper quartile, so approximately 25% of the data lie to the right of this point. This leaves approximately 50% of the data to be covered by the box.

Solution to Activity 18

From Activity 16(a), $q_L = 79$ and $q_U = 162$. So for these data,

$$q_U - q_L = 162 - 79 = 83.$$

That is, the sample interquartile range is 83 pmol/l. (It is good practice to give the units of measures of spread, just as it is for measures of location.)

Solution to Activity 19

The sample variance s^2 is

$$\begin{aligned} s^2 &= \frac{1}{10}\sum_{i=1}^{11}(x_i - 138.6)^2 \\ &= \frac{1}{10}\left\{(66 - 138.6)^2 + (72 - 138.6)^2 + \cdots + (414 - 138.6)^2\right\} \\ &= \frac{1}{10}(5270.76 + 4435.56 + \cdots + 75845.16) = 9598.656. \end{aligned}$$

So the sample standard deviation is $s = \sqrt{9598.656} \simeq 98.0$ pmol/l.

Solution to Activity 20

(a) The sample lower quartile can never be greater than the sample upper quartile, that is, $q_L \leq q_U$, as mentioned towards the end of Subsection 4.2. But the sample interquartile range is $q_U - q_L$, so the interquartile range can never be negative.

(b) From its definition, the sample variance is an average of a set of squared deviations. So as the squared deviations can never be negative, neither can their average. Happily, this means that there should never be the problem of trying to take the square root of a negative number when calculating the sample standard deviation from the sample variance!

(c) The sample standard deviation is just the square root of the sample variance, where the square root is taken to be the positive square root. Hence the sample standard deviation cannot be negative.

Solution to Activity 21

It is striking that in many pairs, the bars for males and females are considerably different in height. This indicates that some occupation types are dominated by females (administrative, caring & leisure and sales) and others are dominated by males (especially skilled trades and operatives, and to a lesser extent managers). Only in the professional, technical and elementary occupation types are the numbers roughly equal.

Solution to Activity 22

(a) Since in this case the area of each bar equals its frequency, the total area is the total frequency, which is $2 + 2 + \cdots + 2 = 49$.

(b) Dividing the frequencies by 49 gives the scaled areas given in the bottom row of Table 15.

Table 15 Completed version of Table 11

Bar number	1	2	3	4	5	6	7	8	9	10	11
Frequency	2	2	2	7	6	7	10	4	5	2	2
Scaled area	0.04	0.04	0.04	0.14	0.12	0.14	0.20	0.08	0.10	0.04	0.04

The sum of the values in the 'Scaled area' row is $0.04 + 0.04 + \cdots + 0.04 = 0.98$. So, after allowing for rounding error, this confirms that the total area of the bars in a histogram using these values is equal to 1.

(c) The scaled areas in Table 15 visually match the heights of the bars in Figure 23 well.

Solution to Activity 23

Both distributions appear to be right-skew and to have prominent modes around 2 seconds. However, the recall times for the unpleasant memories appear to be more spread out than those for the pleasant memories.

Solution to Activity 24

(a) The claim is backed up by Figure 27. The box and the whiskers for the unpleasant memories are longer than the box and the whiskers for the pleasant memories.

(b) The boxplot of the unpleasant memories suggests that the underlying distribution is right-skew. However, the boxplot of the pleasant memories only suggests that the distribution is not symmetric. So the assertion that both distributions are right-skew is only partially backed up by the comparative boxplot.

(c) In Figure 27, it is clear that the unpleasant memories generally took longer to recall. This is because elements of the box for the unpleasant memories are to the right of the equivalent elements of the box for the pleasant memories. There is also a potential outlier in the observations of the pleasant memory recall times but no potential outliers in the observations of the unpleasant memory recall times.

Solution to Activity 25

On this plot, the points generally lie near to a straight line running from bottom left to top right. So there seems to be a positive, linear, relationship between the percentage of adults who were members of a sports club and the percentage of adults who had received coaching. This relationship appears to be fairly strong. There are some potential outliers. For example, the two areas which had the greatest percentages of adult sports club membership had relatively low percentages of adults who had received coaching. Also, one area with adult sports club membership a little over 20% seems to have a noticeably higher percentage of adults who had received coaching than other similar areas.

Solutions to exercises

Solution to Exercise 1

(a) Age in years is a discrete variable, as the numbers will just be whole numbers.

(b) As coded, the length of time the respondent has been living in their current area is an ordinal variable. There are categories, and these categories have a clear ordering.

(c) Marital status is a nominal variable. There are categories, but there is no well-defined ordering of the categories.

(d) How safe the respondent feels is an ordinal variable. There are four categories, and a clear ordering to those categories given by 'degree of safety'.

Solution to Exercise 2

(a) The variables 'how safe the respondent feels' and 'age in years' are linked as they are given as responses by the same set of people.

(b) The variables 'how safe the respondent feels' in the 2014–15 survey and 'how safe the respondent feels' in the 2015–16 survey are not linked as they concern responses by different groups of people.

Solution to Exercise 3

(a) For the dataset not to be a sample, the 22 chondrites would have to represent all the chondrites it is possible to have. This was not so, even at the time of analysis.

(b) A reasonable assumption is that the underlying population is all chondrite meteorites.

(c) While these particular chondrites were certainly not a simple random sample of all possible chondrites, perhaps the best that can be said is that there is no reason to think that the sample is not representative of the population.

Solution to Exercise 4

Much the most common response was for people to say they had been living in their current area for '20+ years'. The least common response was '2–3 years'. The frequency of responses increases from '2–3 years' to '20+ years' and, a little, from '2–3 years' to 'less than 1 year'.

Solution to Exercise 5

The histogram has just a single peak so the distribution appears to be unimodal. This peak is situated a little below zero. Positive values (on the right) appear to be more spread out than negative values (on the left). So the distribution appears to be right-skew: more respondents had a high level of worry about becoming a victim of personal crime than had a low level of worry. Finally, none of the values were less than -2.5 or greater than 3.5.

Solution to Exercise 6

The magnitudes ranged between 2.5 and approximately 5.3, with 50% of them between about 2.6 and 4.6. Overall, the distribution of the magnitudes is right-skew.

Solution to Exercise 7

(a) Since $n = 22$,

$$q_L = x_{\left(\frac{1}{4}(n+1)\right)} = x_{\left(\frac{23}{4}\right)} = x_{\left(5\frac{3}{4}\right)}$$
$$= x_{(5)} + \frac{3}{4}\left(x_{(6)} - x_{(5)}\right) = 26.39 + \frac{3}{4}(27.08 - 26.39) = 26.9075$$

and

$$q_U = x_{\left(\frac{3}{4}(n+1)\right)} = x_{\left(\frac{69}{4}\right)} = x_{\left(17\frac{1}{4}\right)}$$
$$= x_{(17)} + \frac{1}{4}\left(x_{(18)} - x_{(17)}\right) = 33.28 + \frac{1}{4}(33.40 - 33.28) = 33.31.$$

So the sample lower and upper quartiles are approximately 26.91% and exactly 33.31%, respectively.

(b) The sample interquartile range is therefore

$$q_U - q_L = 33.31 - 26.9075 = 6.4025.$$

That is, the sample interquartile range is approximately 6.40%.

Solution to Exercise 8

(a) The ordered data are

$$x_{(1)} = 1, \quad x_{(2)} = 3, \quad x_{(3)} = 4, \quad x_{(4)} = 9, \quad x_{(5)} = 15,$$
$$x_{(6)} = 23, \quad x_{(7)} = 99.$$

Since $n = 7$,

$$m = x_{\left(\frac{1}{2}(n+1)\right)} = x_{\left(\frac{8}{2}\right)} = x_{(4)} = 9,$$
$$q_L = x_{\left(\frac{1}{4}(n+1)\right)} = x_{\left(\frac{8}{4}\right)} = x_{(2)} = 3,$$
$$q_U = x_{\left(\frac{3}{4}(n+1)\right)} = x_{\left(\frac{24}{4}\right)} = x_{(6)} = 23$$

and

$$q_U - q_L = 23 - 3 = 20.$$

(b) $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{99 + 23 + 1 + 3 + 15 + 4 + 9}{7} = \frac{154}{7} = 22,$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{6}\sum_{i=1}^{7}(x_i - 22)^2$$
$$= \frac{1}{6}\left\{(99 - 22)^2 + (23 - 22)^2 + \cdots + (9 - 22)^2\right\}$$
$$= \frac{1}{6}(5929 + 1 + \cdots + 169) = \frac{7274}{6} \simeq 1212.3$$

and

$$s = \sqrt{\frac{7274}{6}} \simeq 34.8.$$

(You should have obtained precisely the same results in this part had you worked with the ordered rather than the unordered data. Sample means and standard deviations are not changed by reordering.)

(c) The median number of operational nuclear power stations and the mean number of operational nuclear power stations are rather different, at 9 and 22, respectively. This difference makes it hard to state a 'typical' number of operational nuclear power stations, an effect caused by the unusually large data value, 99, included in the sample. The median is usually preferred in such a situation, so having nine operational power stations might be considered typical. The same extreme value is responsible for the sample standard deviation (34.8) being very large. The resistant sample interquartile range also takes a rather large value, namely 20. There seems to be a high degree of spread across these data values; the number of operational nuclear power stations varies widely from country to country.

Solution to Exercise 9

The percentages of deaths attributable to high blood glucose were higher in low-income and lower-middle-income countries than in the richer countries – and this was the same for men and women. However, for countries with the same general income levels, there were lower percentages of deaths attributable to high blood glucose in females compared with males.

Solution to Exercise 10

(a) In both of the figures, it is clear that the number of children is typically lower for the mothers that spent longer in education.

(b) In the unit-area histograms, it is clear that the distribution of the number of children in the group corresponding to the mothers with the shorter education is bimodal, whereas the distribution appears to be unimodal for mothers who had a longer education. This cannot be seen from the comparative boxplot.

(c) The comparative boxplot suggests that in the group of mothers who had a longer education, a few had unusually large numbers of children. This feature is less obvious (though present) in the unit-area histograms.

Solution to Exercise 11

The scatterplot suggests that there is no clear-cut relationship between the magnitude and depth of an earthquake, though the deeper earthquakes do seem to have bigger magnitudes than most of the shallower ones.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Page 3: © Kran Kanthawong / 123RF.com

Page 4: Avda / https://commons.wikimedia.org/wiki/File:Kernkraftwerk_Grafenrheinfeld_-_2013.jpg This file is licensed under the Creative Commons Attribution-ShareAlike Licence http://creativecommons.org/licenses/by-sa/3.0/

Page 5 top: © iStockphoto.com / Highwaystarz-Photography

Page 5 bottom: Taken from: Canada. Dept. of Interior / Library and Archives Canada / C-036149

Page 6 top: © Patrick Callaghan

Page 6 bottom: Taken from: http://www.wikihow.com/Avoid-the-Temptation-to-Eat-Unhealthy-Foods

Page 7: © damiangretka / www.istockphoto.com

Page 8: © Dustin M. Ramsey

Page 9: © Copyright NewcastleGateshead

Page 12: Vladimir Seliverstov / www.123rf.com

Page 14: H. Raab This file is licensed under the Creative Commons Attribution-ShareAlike Licence http://creativecommons.org/licenses/by-sa/3.0/

Page 17: © Yauheniya Hauss / 123RF.com

Page 19: © Gregg from GriDD / www.whatthegregg.com

Page 23: Taken from: https://twitter.com/hashtag/foodpackaging?src=hash

Page 26 top: © KGalione / www.istock.com

Page 26 bottom: © Richard Wong / Alamy Stock Photo

Page 27: © Melissa Nemeth This file is licensed under the Creative Commons Attribution-ShareAlike Licence https://creativecommons.org/licenses/by-sa/2.0/

Page 29: © Hulton Archive / Stringer / Getty Images

Page 31: © Ben Chun / https://flic.kr/p/baSfHP This file is licensed under the Creative Commons Attribution-ShareAlike Licence https://creativecommons.org/licenses/by-sa/2.0/

Page 34: © Jakec / https://en.wikipedia.org/wiki/File:Interstate_80_in_Hemlock_Township,_Columbia_County,_Pennsylvania.JPG This file is licensed under the Creative Commons Attribution-ShareAlike 3.0 Unported License

Page 38: © Ian Britton / http://www.freefoto.com/preview/41-14-12/Dual-Carriageway This file is licensed under the Creative Commons Attribution-NonCommercial-NoDerivative 3.0 License

Page 39: Steve Karg / https://commons.wikimedia.org/wiki/File:Western-pack-butter.jpg This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Page 40: © Philip Bird / 123RF.com

Page 42: © Meme Center

Page 45: © kzenon / 123RF.com

Page 49: © David Mzareulyan / www.istockphoto.com

Page 53: © flairmicro / www.123rf.com

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked, the publishers will be pleased to make the necessary arrangements at the first opportunity.